From the Department of Neuroscience
Karolinska Institutet, Stockholm, Sweden

GENETIC ANALYSIS AND SOFTWARE DEVELOPMENT FOR STUDIES ON COMPLEX HUMAN TRAITS

Lisette Graae

Karolinska Institutet

Stockholm 2014

Cover picture: Our knowledge of the human genome today probably only represents the tip of an iceberg. Getting to the bottom will likely require sophisticated methods and a lot of endurance from the scientific community.

Author kayaking at Greenland summer 2003. Photo taken by Christopher Graae.

Published by Karolinska Institutet. Printed by Larserics Digital Print AB.

ISBN 978-91-7549-365-7

Institutionen för neurovetenskap

Genetic analysis and software development for studies on complex human traits

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Hillarpsalen, Retzius väg 8

Fredagen den 10 januari, 2014, kl 09.00

av

Lisette Graae

Huvudhandledare:
Docent Andrea Carmine Belin
Karolinska Institutet
Institutionen för neurovetenskap

Bihandledare:
Docent Silvia Paddock
Rose Li and Associates, Inc
Bethesda, MD, USA

Docent Dagmar Galter
Karolinska Institutet
Institutionen för neurovetenskap

PhD. Maria Lindskog
Karolinska Institutet
Institutionen för neurovetenskap

Fakultetsopponent:
Docent Lars Feuk
Uppsala Universitet
Institutionen för immunologi, genetik och patologi

Betygsnämnd:
Professor Catharina Larsson
Karolinska Institutet
Institutionen för onkologi-patologi

Docent Erik Jönsson
Karolinska Institutet
Institutionen för klinisk neurovetenskap

Professor Niklas Dahl
Uppsala Universitet
Institutionen för immunologi, genetik och patologi

Stockholm 2014

ABSTRACT

Rapidly advancing genotyping and genome sequencing technologies have recently given us the opportunity to watch evolution at work. It turns out that the genome, like so many other workplaces, is rather messy. The sequences of no two humans are exactly alike, not even those of "identical" twins. A lot of this variability has allowed us as a species to adapt to different living conditions. But some of it leads to disease. This thesis consists of four studies aimed at elucidating the genetic mechanisms involved in heritable complex genetic disorders of the brain. In **paper I**, we studied the involvement of genome-wide sequence variants of estrogen receptor (ER) binding sites in mood disorders. Based on previously reported gender differences we studied each gender separately and found that a polymorphism in the genetic code close to the gene *TGM2* might be involved in the disease mechanism for bipolar disorder (BD) affecting women. During this study we developed several independent Perl programs, which could be useful also to other genetic researchers. Therefore, in **paper III**, we merged these programs together into one user-friendly tool called ReMo-SNPs and extended its features so it would be able to perform searches in any region or motif of interest genome-wide. By combining *in silico* identified binding motifs with experimentally validated *in vivo* or *in vitro* binding regions, the program enables the researcher to maximize the functionality and relevance of the identified SNPs. In **paper II** we were interested in evaluating the possible involvement of copy number variations (CNVs) in BD, and schizophrenia (SZ). Based on linkage studies, we selected 227 individuals with high disease prevalence for genome-wide genotyping. CNV variation was assessed and a ~200 kilobase (kb) deletion in the *MAGI1* gene was identified. Further analyses of CNVs in the *MAGI1* and *MAGI2* genes were performed in a pooled analysis comprising 10,925 BD or SZ cases and 16,747 controls. We identified eleven additional CNVs of >100 kb in the *MAGI1* and the *MAGI2* genes in cases, while only four CNVs were found in these genomic regions in the control group. An association analysis on the pooled sample resulted in a significant association (p-value 0.023). This supports a possible role for rare *MAGI1* and *MAGI2* CNVs as risk factors for BD and SZ. In **paper IV** we tested associations of eight variations previously identified in GWAS in large European cohorts of migraine patients in a Swedish material. We found support for two of the previously identified SNPs, rs1835740 and rs2651899, which means that they might be involved in the disease mechanism also in Sweden. **Conclusions**: The work performed in this thesis has increased our knowledge of the genetic mechanisms behind BD, SZ, major depression, and migraine, four complex genetic disorders. A SNP in an ER binding region close to the gene *TGM2* might be involved in the disease mechanism for women with BD (**paper I**). Large (>100 kb), rare CNVs in the genes *MAGI1* and *MAGI2* might be involved in the disease mechanisms in BD and SZ (**paper II**). Our ReMo-SNPs program has become a useful tool to identify functional SNPs for association studies in any region or motif of interest genome-wide (**paper III**). Two SNPs, rs183740 and rs2651899, might be involved in the disease mechanism in Swedish patients with migraine (**paper IV**). Even though our findings only constitute a small piece in the gigantic puzzle of complex genetic disorders, they may be of help to future researchers, clinicians, and patients and bring us a few steps forward in our quest to sort out the messy workplace that is our genome.

LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals:

I. **Lisette Graae**, Robert Karlsson, and Silvia Paddock

Significant Association of Estrogen Receptor Binding Site Variation with Bipolar Disorder in Females

PLoS ONE 7(2): 2012; e32304. doi:10.1371/journal.pone.0032304

II. Robert Karlsson, **Lisette Graae**, Magnus Lekman, Dai Wang, Reyna Favis, Tomas Axelsson, Dagmar Galter, Andrea Carmine Belin, and Silvia Paddock

MAGI1 Copy Number Variation in Bipolar Affective Disorder and Schizophrenia

Biological Psychiatry. 2012; 71(10):922-930

III. **Lisette Graae**, Andrea Carmine Belin, and Silvia Paddock

ReMo-SNPs: A new software tool for identification of polymorphisms in regions and motifs genome-wide

Manuscript

IV. Caroline Ran, **Lisette Graae**, Patrik K.E. Magnusson, Nancy L. Pedersen, Lars Olson, and Andrea Carmine Belin

Genetic studies on sporadic migraine in a Swedish cohort

Manuscript

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AMPA	α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor
ANK3	ankyrin-3
BD	bipolar disorder
CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit
CACNB2	calcium channel, voltage-dependent, beta 2 subunit
CEU	Utah residents with Northern and Western European ancestry from the Centre d'Etude du Polymorphisme Humain collection
ChIP	chromatin immunoprecipitation
CI	confidence interval
CIDI	Composite International Diagnostic Interview
CNV	copy number variation
dbGaP	Database on Genotypes and Phenotypes
dbSNP	Database of Single Nucleotide Polymorphsims
DGV	Database of Genomic Variants
DIGS	Diagnostic Interview for Genetic Studies
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DSM	Diagnostic and Statistical Manual of Mental Disorders
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait loci
ER	estrogen receptor
ErbB4	erythroblastic leukemia viral oncogene homolog 4
FHM	familial hemiplegic migraine
GR	glucocorticoid receptor
GWAS	genome-wide association study
HWE	Hardy-Weinberg equilibrium
ICD	International Classification of Diseases

ICHD	International Classification of Headache Disorders
LD	linkage disequilibrium
lincRNAs	long intergenic non-coding RNAs
LOD	logarithm of the odds
MAF	minor allele frequency
MAGI	membrane-associated guanylate kinase
MDS	multidimensional scaling
MHC	major histocompatibility complex
MIR137	microRNA 137
MTDH	metadherin
NCAN	neurocan
NCBI	National Center for Biotechnology Information
NESDA	Netherlands Study of Depression and Anxiety
NHGRI	National Human Genome Research Institute
NIMH	National Institute of Mental Health
NTR	Netherlands Twin Register
ODZ4	a human homologue of the *Drosophila* pair-rule gene *ten-m*
OR	odds ratio
PCR	polymerase chain reaction
PPAR	peroxisome proliferator-activated receptor
PRDM16	PR domain containing 16
QC	quality control
RNA	ribonucleic acid
SA	schizoaffective disorder
SLC1A2	solute carrier family 1, member 2
SNP	single nucleotide polymorphism
SZ	schizophrenia
TF	transcription factor
TGM2	transglutaminase 2

WHO	World Health Organization

1 INTRODUCTION

This thesis comprises four studies (papers I-IV) on genetic variation in complex psychiatric and neurological diseases: bipolar disorder, major depression, schizophrenia, and migraine. The first section of the thesis briefly describes the key genetic concepts upon which these studies rest, followed by a description of the four complex diseases and the state of the art regarding the identification of their genetic components.

1.1 Complex genetics

Monogenic traits result from a single genetic mutation causing the trait. Most human genetic traits are, however, caused by the combination of many genes and are said to be multigenic. Research on complex human diseases in the last two decades indicates that they usually involve complex, multigenic risk factors in combination with environmental influences. The contribution of each gene or mutation to the trait is often small and may therefore be difficult to identify in genetic studies. A disease with a complex genetic background does not follow Mendelian patterns of inheritance, even though these diseases often aggregate in families to some extent. The inherited genes associated with the disease often only infer a predisposition to develop the disorder and other factors (e.g., life-style, and environmental factors) are also involved in the disease etiology.

1.2 Genomic variability

The human genome constitutes of approximately 3 billion nucleotide base pairs elegantly stored in the DNA structure. There is not just one human genome sequence, because all individuals harbor a unique combination of nucleotides in their genome. The variable sequences are, however, largely limited to a tiny part that differs between us (about 0.1% of the total sequence length). Variations in these variable genomic positions, which are called polymorphisms, make each individual unique. These genomic polymorphisms in combination with different environmental exposures cause people to look different, respond differently to medical treatment, and develop different diseases or not. There are several types of genetic variations. Some alterations are large and include whole or part of chromosomes whereas other only involve a change in a single nucleotide base pair.

1.3 Single nucleotide polymorphisms (SNPs)

SNPs are the most common type of genetic variation in the human genome, with more than 60 million SNP variants reported for the human genome in the National Center for Biotechnology Information (NCBI) Database of Single Nucleotide Polymorphsims (dbSNP) Build 138 (Sherry et al., 2001). A SNP is a DNA variation in a single nucleotide base; for example when a cytosine nucleotide (C) is replaced with a thymine nucleotide (T) in the DNA sequence. Such a nucleotide substitution creates two possible alleles in that position of the genome, the ancestral allele C and the newly formed T-allele. Because human genomes are diploid, meaning that we have two copies of each chromosome,

these alleles can be inherited in three different combinations (CC, CT and TT). Allele frequencies can vary between different populations and different ethnical groups as well as between sick and healthy individuals for a certain trait. The minor allele frequency (MAF) specifies the frequency of the less common allele for the population studied.

Approximately 80% of all SNPs are present in all major population groups, since most of the variation found in the world today originates from a time before humans dispersed out of Africa (Jakobsson et al., 2008). However, there is greater genetic variation in the African population compared to other populations in the world, because only a small subgroup of all people migrated from Africa (Jorde et al., 2001).

SNPs may be located in coding or non-coding parts of the genome. A SNP in a coding region can either be synonymous or non-synonymous, where a non-synonymous SNP causes a change in the amino acid sequence that constitutes the backbone of the translated protein. SNPs outside coding parts of the genome may still exert an effect on the protein expression by for example affecting splice sites or regions that are involved in the regulation of the amounts of proteins made.

1.4 Structural variation in the genome

SNPs are the most common type of genetic variation in the human genome, but when measuring the difference between individuals in genetic base pairs, structural changes contribute most, because they may encompass millions of bases of DNA and often include entire genes and their regulatory regions. A structural variation is defined as any variation larger than a SNP or micro-satellite variation, which are blocks of usually two to six base pairs repeated ten to hundreds of times (Wain et al., 2009). Some structural variants are large (~3Mb or more) and include changes of whole or parts of chromosomes, and can be made visible under a microscope. Other changes are smaller and can be detected by more sensitive methods like sequencing, as well as genomic microarrays including SNP- and genomic hybridization arrays. There are several different types of structural variations; some are balanced changes in DNA content, (e.g, inversions, balanced translocations), which rearrange the DNA content without adding or removing sequence information. Copy number variation (CNV) on the other hand is a type of un-balanced DNA variation that results in a gain or loss of DNA information by deletion, duplication or insertion of DNA sequences into the genome. CNVs usually refer to DNA sequences of about 1000 or more base pairs in length (Wain et al., 2009). They have been found to be ubiquitous and affect up to 12% of the human genome (Feuk L., 2012). Each individual genome harbors hundreds of CNVs, but most CNVs are observed only rarely (<1%) in the human population (Johansson and Feuk, 2011).

Copy number variants thus account for a substantial amount of genetic variation and play an important role for individual variation, evolution and disease development. Even monozygotic twins have been shown to have slightly different CNV distributions within their genomes (Bruder et al., 2008). A tremendous advancement in the technologies and informatics tools to discover and interpret CNVs during the last 10 years has led to rapid

advances in our understanding of the nature of CNVs in the human genome. The Database of Genomic Variants (DGV) today provides a catalog of data on CNVs in healthy control individuals based on 55 published studies, comprising >2.5 million entries identified in >22,300 genomes (MacDonald et al., 2013).

Several studies during the past few years have focused on characterizing disease-causing CNVs, whereas others have tried to establish a map of the normal CNV variation in genomes of healthy individuals. A lot remains to be understood regarding the CNV distribution in the genomes of healthy individuals as well as in individuals with various disorders.

1.5 Recombination and linkage disequilibrium

During cell divisions of gametes (sperm and egg cells), pairs of paternal and maternal homologous chromosomes line up beside each other and recombination (crossing over) occurs between them. This ensures that each gamete contains a new unique genetic combination including both maternally and paternally derived genetic information, which is passed on to the next generation. This crossing over of genetic material can split alleles that lie next to each other on a chromosome and lead to the formation of new haplotypes (the combination of alleles along a chromosome). The recombination fraction measures genetic distance by describing the extent of linkage between two loci. The further apart two loci are situated the higher the probability that a recombination will occur between them, and the closer two loci are on the same chromosome the greater their chance of being inherited together. There is, however, no absolute correlation between physical and genetic distance. Genetic distance is measured in centimorgans, where 1 centimorgan roughly corresponds to 1 billion bases (Burton et al., 2005).

Since alleles close to each other tend to be inherited together, there is a non-random association between alleles in a population, a phenomenon that is called linkage disequilibrium. This means that certain combinations of alleles are observed more often than would be expected if they were completely unlinked. Linkage disequilibrium implies that the information given by one allele also provides information about other alleles. SNPs in linkage disequilibrium with each other are said to be located in a haplotype block. By selecting some of the SNPs from these haplotype blocks, one can get information on a majority of the SNP variation within each block. This mechanism is utilized in both linkage and association analyses where some of the common SNPs (or microsatellite markers) in these haplotype blocks are selected for the genotyping platforms (Cordell and Clayton, 2005; Dawn Teare and Barrett, 2005). Linkage disequilibrium can be influenced by many different factors including genetic drift, natural selection, as well as population growth and structure (Palmer and Cardon, 2005).

1.6 Linkage analyses

Genetic linkage analyses aim to map genetic loci to certain traits in related individuals. By studying the transmission of markers in a pedigree with multiple affected family

members, one can trace which markers and hence genomic regions co-segregate with the disease. It is a method to identify broad genomic regions that often contain several genes. Fine-mapping of the most interesting regions is often performed as a second stage in these analyses. There are two main types of linkage analyses: parametric linkage analysis, which is a model-based analysis usually applied to monogenetic disorders, and non-parametric analysis, which is a model-free analysis used for complex genetic disorders (Dawn Teare and Barrett, 2005). The degree of linkage is usually reported as a logarithm of the odds (LOD) score, where large positive scores are evidence for linkage and negative scores are evidence against. Results from non-parametric studies are reported as NPL scores, which are a form of a Z-score. Linkage analyses works best in diseases where the underlying mechanisms are caused by only a few highly penetrant variants.

1.7 Association analyses

In association analyses one studies the patterns of polymorphisms that vary systematically between individuals with different disease states (e.g., cases and controls) in order to identify the alleles that increase risk or protect from disease. Association analyses can be considered a special form of linkage analyses based on the fact that the human population shares common ancestry and thus could be viewed as an extended family (Cordell and Clayton, 2005). Association analyses have greater power than linkage studies to detect common variants with a modest or small effect size and have become a popular tool for studies on complex genetic diseases (Risch and Merikangas, 1996). An association study starts with collection of biological samples from the study participants, followed by DNA extraction and genotyping of the selected markers. After several data quality control steps the frequency of each allele is calculated and compared between, for example, the groups of cases and controls. A large difference in genotype frequency could indicate that the allele might be involved in the disease or is in linkage disequilibrium with a marker involved in the disease.

There are several types of association studies: 1) the candidate polymorphism study, which focuses on one polymorphism at a specific locus; 2) the candidate gene approach, which focuses on previously identified genes (either from linkage studies or functional studies) related to the disease and usually involves 5-50 SNPs; 3) the fine mapping of one or several interesting regions previously identified as related to the disease (usually by linkage studies), which involves several hundred SNPs; and 4) the whole-genome association study (GWAS), which is a hypothesis-free approach with several hundred thousand up to a million SNPs dispersed throughout the genome. Recent technical advancements have reduced the costs for GWAS remarkably, making this approach to association studies increasingly popular. The National Human Genome Research Institute (NHGRI) has established a catalog of published GWAS (www.genome.gov/gwastudies) (Hindorff et al., 2009). At the last update (11/20/13) it included 1,757 publications with 11,953 SNP-disease associations reported with a significant threshold of $p< 1x10^{-5}$. Out of these, however, only 34 involve bipolar disorder; 39 schizophrenia; 5 depression; and 6 migraine, which clearly indicates that there is a

great need of more research in these disorders. Figure 1 illustrates the amount of published findings with a p value of $1x10^{-5}$ or better from GWAS in schizophrenia (39) in comparison with cancer-GWAS (152).

A complicating consequence of the genome-wide design is the problem with multiple testing. Since each SNP represents one test, an enormous amount of tests is performed, which makes it important to apply stringent significant thresholds to reduce the amount of false positive, type I errors. An often used significant threshold for GWAS are p-values equal to or less than 5×10^{-8}, which takes into account a total number of 1 million SNPs tested. To obtain this kind of significance, one needs very large sample sizes just to overcome the multiple testing problem. Furthermore, effect sizes in complex genetic disorders are usually very small, further increasing the required number of study participants. A popular approach has therefore become to do a multi-stage analysis including a discovery set of individuals that undergo a GWAS. Interesting findings from this initial study are then followed up in a second stage, where the significance thresholds are lower, because fewer markers are tested. Another way to gain large enough sample sizes is through mega- and meta-analyses. Such collaboration efforts have become increasingly popular, and large consortia have been formed for many diseases, such as the International Headache Genetics Consortium (http://www.headachegenetics.org/) involved in research on headache and related disorders, and the Psychiatric Genetics Consortium (http://pgc.unc.edu) focusing on genetic studies on autism, attention deficit hyperactivity disorder, bipolar disorder, schizophrenia, and major depression.

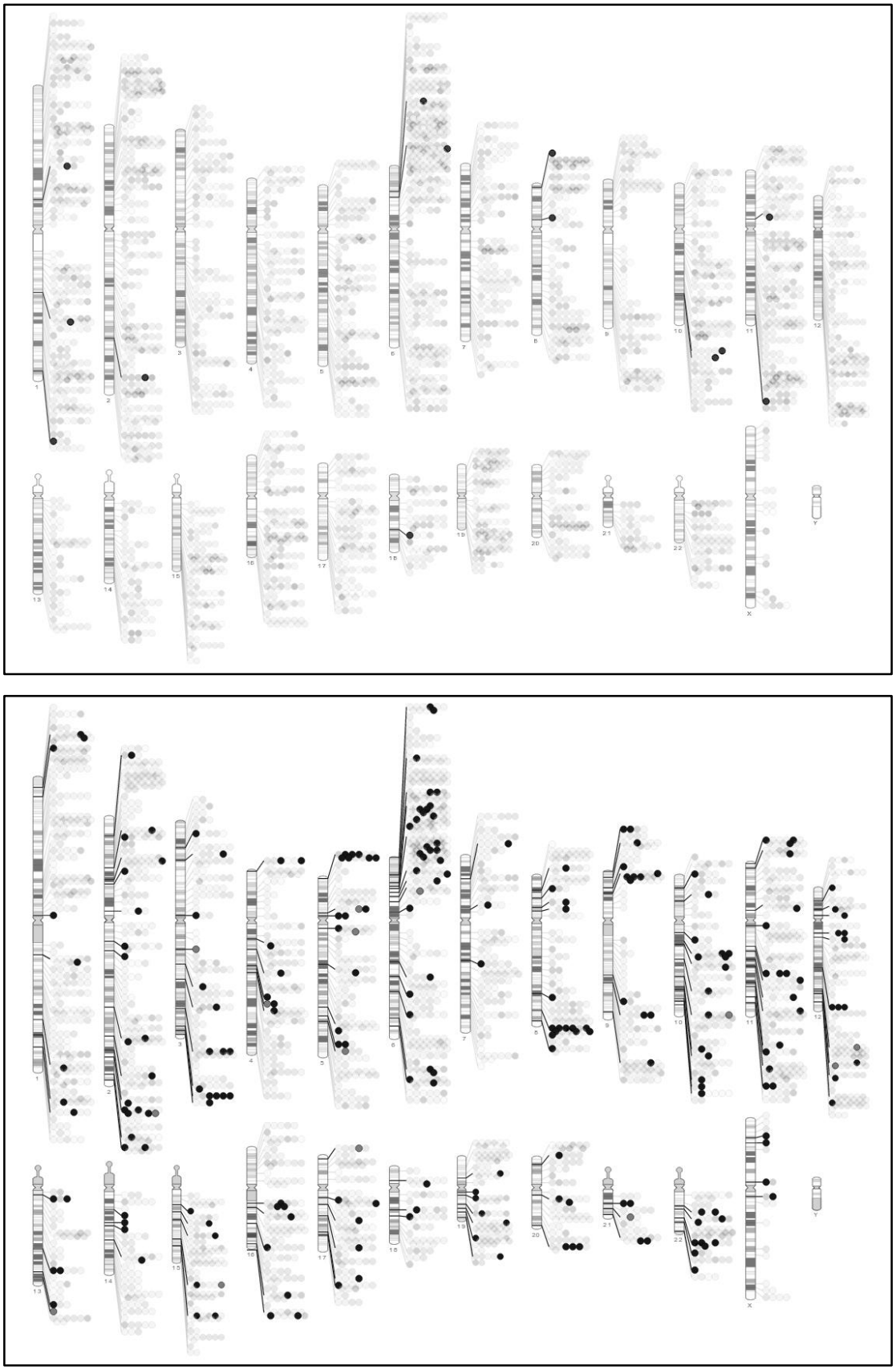

Figure 1 Published GWAS findings with a p-value of $1x10^{-5}$ or better as reported in the catalog of published GWAS made by NHGRI (11/20/13). Reported GWAS findings illustrated as filled circles for each chromosome. Top picture shows GWAS findings in schizophrenia (n=39) and the bottom picture in cancer (n=152).

1.8 Transcription factors (TFs)

Transcription factors are proteins that regulate gene transcription by binding to specific DNA sequences, such as promoters, enhancers, silencers, insulators and locus control regions (Yip et al., 2012). In humans, approximately 1850 different TFs have been identified that are involved in gene regulation through processes of activation or repressing the transcription (Maston et al., 2006). TFs not only bind to the vicinity of the genes they regulate, but they are also involved in gene regulation from a distance, sometimes through influencing modifications of the chromatin structure and thereby the accessibility of the transcription machinery (Yip et al., 2012). TFs rarely operate by themselves but rather form complexes with several other TFs and DNA-binding proteins (Blanchette, 2006). The Encyclopedia of DNA Elements (ENCODE) is a project started by NHGRI in 2003 with the aim to delineate all functional elements encoded in the human genome (Consortium, 2012). One of their goals is to create a genome-wide map on transcription factor binding regions in diverse human cell types identified through high-throughput ChIP-seq *in vivo* (The ENCODE Project Consortium, 2011). So far, they have found that the fraction of bases in the genome likely to be involved in gene regulation is significantly higher than the number of bases found in protein coding exons, which indicates that more information in the human genome may be important for gene regulation than for biochemical function (Consortium, 2012). This notion is also supported by the fact that 88% of the SNPs identified as associated with a phenotype in recent GWAS results are placed in either intronic or intergenic genomic regions. 12% of these SNPs are found to overlap with TF binding regions, and when also taking into account SNPs in linkage disequilibrium, as much as 31% of the identified GWAS loci have a SNP that overlaps a TF-binding site (Consortium, 2012). This indicates the importance of further studies in these regions (e.g. by candidate region association analyses), in order to elucidate pathways and mechanisms that may underlie complex genetic disorders.

2 THE DISEASES

The studies in this thesis focus on three complex psychiatric diseases (bipolar disorder, major depression, and schizophrenia) and one neurological disorder (migraine). All four traits have been shown to be highly heritable with a complex mode of inheritance involving several genes and possible also environmental factors. Currently no known major disease loci or biological markers have been identified, and the diagnosis is mainly based on self-reports and clinical observations. Pharmacological treatments for the psychiatric disorders often involve several side effects and sometimes, as for schizophrenia, not all symptoms are treatable with medicine. The diseases often result in life-long suffering for the patients and their relatives, and the economic burden to society is considerable.

2.1 Diagnostic manuals

For the psychiatric disorders there are two diagnostic manuals: the International Classification of Diseases (ICD) published by the World Health Organization (WHO) and the Diagnostic and Statistical Manual of Mental Disorders (DSM) produced by the American Psychiatric Association (American Psychiatric Association, 2013; World Health Organization (WHO), 1994). The classification systems correlate highly with each other. The DSM is more often used in US-based cohorts, while clinicians in Europe use the ICD more frequently for their studies. The latest version of the ICD, ICD-10, came into use in 1994. A new edition of the classification manual will be released in 2015. The fifth edition of the DSM was released in May 2013.

Diagnostic criteria for migraine have been established by the International Classification of Headache Disorders (ICHD-II)(Headache Classification Subcommittee of the International Headache Society, 2004). The beta-version of the 3rd edition has recently been released (Headache Classification Committee of the International Headache Society, 2013).

2.2 Bipolar disorder

Symptoms

Bipolar disorder is an episodic mood disturbance that lets the affected individuals experience extreme shifts in mood ranging from elation or mania to severe depression, sometimes accompanied by disturbances in thinking and behavior. Psychotic features such as delusions and hallucinations are frequently encountered. According to DSM-5, there are three major subtypes of bipolar disorder: bipolar I disorder, bipolar II disorder, and cyclothymic disorder. A bipolar I diagnosis requires an experience of a period of mania, with or without previous experience of a hypomanic or major depressive episode. During a manic episode the patients experience abnormally and persistently elevated mood and increased goal-directed activity or energy. An inflated self-esteem in combination with poor judgment often lead to activities with painful consequences, such

as giving away possessions, foolish business investments, and sexual promiscuity. The diagnostic manuals require that the symptoms last for at least one week and be severe enough to impair social or occupational functioning or even require hospitalization. A hypomanic period is described as a milder form of mania (both regarding duration and symptoms) and commonly occurs in bipolar disorder type I and type II. To fulfill criteria for a bipolar II diagnosis, the patient should have experienced a hypomanic and a major depressive episode (classification criteria for a major depressive episode are described in the next section). Cyclothymic disorder patients never fulfill the criteria for an episode of mania, hypomania or major depression, but experience similar symptoms of mood swings for at least two years. Age at onset is usually between late adolescence and the mid-20s. Females are more likely to experience depressive symptoms and to have a rapid cycling mood or experience mixed states.

Genetics

Genetic factors have long been suggested to be an important contributor to the disease etiology for bipolar disorder. Lifetime risk calculations for individuals in the general population show a risk of 0.5-1.5% of getting ill, whereas for a first degree relative it is 5-10%, and for monozygotic twins it is as high as 40-70% (Craddock and Sklar, 2013). The fact that there is not a 100% concordance between monozygotic twins suggests that other factors than genes play an important role in the disease mechanisms.

It has become evident that genetic risk factors involved in bipolar disorder likely imply the contribution of many (possible >100) common nucleotide changes with low individual effects and possible also rare CN changes. The genetic mechanisms are likely complex, with the involvement of and interplay between several genes and pathways (Craddock and Sklar, 2013).

Early linkage and candidate gene association studies did not yield any robust positive findings and showed very little consistency between studies (Barnett and Smoller, 2009), a finding that supports the theory of the involvement of many common variants rather than a few variants with large effects. The positive results seen in some of these studies may, of course, contribute to the disease mechanisms in the individuals included in the studies, but they may not explain the disease pathways for the major part of individuals with the disorder, which could be one reason why replication studies often failed. Another reason is the possibility that many of the individually significant studies report false positive findings. Without hindsight, it is not possible to tell which studies will replicate and which won't, and several complex disease factors were first identified in large meta-analyses of conflicting results.

GWAS on bipolar disorder have been somewhat successful in identifying genes with possible involvement in the disease mechanisms. Several studies have identified markers with genome-wide significance, and some of them have also been replicated in other cohorts. The most replicated finding implies variants in the *CACNA1C* gene, which encodes the alpha 1C subunit of the voltage gated L-type calcium channel. This protein

is involved in learning, memory and synaptic plasticity and is widely distributed in the brain (Ferreira et al., 2008; Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011). Some mood-stabilizing drugs (eg verapamil and diltiazem) regulate ion channels, which further supports a role for this pathway (Craddock and Sklar, 2013). Other genes with genome-wide significance in at least one study are *ANK3* (encoding ankyrin 3), a protein that likely links membrane proteins to the cell cytoskeleton), *NCAN* (encoding neurocan, which modulates cell adhesion and migration), and *ODZ4* (encoding a human homologue of the *Drosophila* pair-rule gene *ten-m*) (Cichon et al., 2011; Ferreira et al., 2008; Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011).

There are fewer studies supporting a role for the involvement of CNVs in bipolar disorder compared to schizophrenia. One genome-wide study reported higher frequencies of *de novo* CNVs in bipolar disorder cases compared to controls, especially for cases with an early onset of the disease (Malhotra et al., 2011).

2.3 Major depression

Symptoms

There are several types of depressive disorders that all share the features of a sad, empty, or irritable mood with additional somatic and cognitive symptoms that significantly influence the individual's capacity to function. Major depressive disorder is a recurrent mood disorder with episodes lasting at least two weeks but often longer. Symptoms characterizing a major depression episode are: depressed mood, loss of interest or pleasure in activities, change in appetite, insomnia or hypersomnia, psychomotor agitation or retardation, fatigue or loss of energy, feelings of worthlessness, diminished ability to concentrate, and recurrent thoughts of death. Accompanying the symptoms are impairments in social, occupational, or other important areas of functioning.

Reported prevalence of major depression ranges globally between 0.8 percent and 10.4 percent, depending on study methodology and sampling (Ferrari et al., 2013). The prevalence reported for women is 1.5 to 3-fold higher compared to men (Kessler et al., 2003). Major depressive disorder may appear at any age throughout life, but there is a peak in onset in the 20s (American Psychiatric Association, 2013).

Genetics

Compared to genetic studies in other psychiatric diseases, studies aiming at revealing the genetic risk factors involved in major depression have not been as fruitful. None of the nine large published GWAS so far have reported any robust loci with genome-wide significance (Kohli et al., 2011; Lewis et al., 2010; Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium et al., 2013; Muglia et al., 2010; Rietschel et al., 2010; Shi et al., 2011; Shyn et al., 2011; Sullivan et al., 2009; Wray et al., 2012). This may be a bit surprising considering the large number of individuals included in the

studies. The most recent study published by the Psychiatric GWAS Consortium included 9240 major depression patients and 9519 controls. These results indicate that the genetic contribution to major depression likely constitutes of a rather large number of common variants with small individual effects.

There are surprisingly few studies on the contribution of structural variations to the underlying genetic mechanisms of major depression. One recent GWAS, analyzing CNVs in a sample of recurrent depressive disorder, found that rare, large (>100kb) deletions were significantly enriched in cases and that the overall load of deletions was larger in major depression cases compared to controls (Rucker et al., 2013).

2.4 Schizophrenia

Symptoms

Schizophrenia is a mental disorder with symptoms in one or more of the following five areas: delusions, hallucinations, disorganized thinking, grossly disorganized or abnormal motor behavior, and negative symptoms. The DSM-5 defines key features of the disorder. Delusions are fixed beliefs that are not amenable to change in the light of conflicting evidence. Hallucinations are perception-like experiences that occur without an external stimulus, where auditory hallucinations are the most common type. Disorganized thinking typically implies disorganized speech, and the symptoms must be severe enough to substantially impair effective communication. Grossly disorganized or abnormal motor behavior can range from childlike behavior to unpredictable agitation and often leads to difficulties in performing activities of daily living. Catatonic behavior, defined as a marked decrease in reactivity to the environment, is also included in this category of symptoms. The negative symptoms account for a substantial part of the schizophrenia morbidity; they are more closely related to prognosis than the positive symptoms and tend to be more persistent. The two most prominent negative symptoms in schizophrenia are diminished emotional expression and avolition, which means an inability to initiate or continue with goal-directed activities.

Several sub-classifications of schizophrenia exist that are not described in detail herein. In addition to cases diagnosed with bipolar disorder and schizophrenia, in paper II we also included individuals diagnosed with schizoaffective disorder, which can be considered an intermediate form of the two disorders, where the diagnosis depends on the number, severity and duration of the schizophrenic and affective mood symptoms.

Depending on the diagnostic criteria used (from narrow to broad), the reported prevalence for schizophrenia varies between 0.41 percent and 1.54 percent (Lichtenstein et al., 2006). Compared to women, men often experience a more severe form of schizophrenia with more pronounced negative symptoms, earlier onset of the disease, and longer duration of illness (van Os and Kapur, 22). A majority of individuals gradually develop symptoms associated with schizophrenia. The peak in onset for the first psychotic episode occurs in the early to late 20s (McGrath et al., 2008).

Genetics

Twin studies indicate a strong genetic component to the disease mechanism underlying schizophrenia with heritability estimates of around 80 percent (van Os and Kapur, 22). In a recent multi-stage GWAS on schizophrenia, 22 loci harboring common variants were identified as associated with the disease (Ripke et al., 2013). The authors speculate that probably between 6,300 and 10,200 independent and mostly common SNPs contribute to the etiology of schizophrenia. These SNPs are, however, thought to be clustered in a limited number of biological pathways important to the disease pathology. The SNPs with a genome-wide significance in the most recent were clustered in four biological pathways or regions that may be implicated in the disease etiology. These pathways or regions involved 1) calcium signaling; 2) the major histocompatibility complex (MHC); 3) microRNA 137 (MIR137); and 4) long intergenic non-coding RNAs (lincRNAs). Two SNPs placed in genes coding for voltage-dependent L-type calcium channel subunits (*CACNA1C* and calcium channel, voltage-dependent, beta 2 subunit *(CACNB2)*) were also identified. These genes have previously been identified both in disease-specific GWAS (bipolar disorder, schizophrenia, and major depression) as well as in a meta-analysis looking at shared risk loci between five psychiatric disorders (bipolar disorder, schizophrenia, major depression, autism spectrum disorder, and attention deficit-hyperactivity disorder) (Cross-Disorder Group of the Psychiatric Genomics Consortium and Genetic Risk Outcome of Psychosis (GROUP) Consortium, 2013; Ferreira et al., 2008; Green et al., 2010; Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011). Variants in the *CACNA1C* gene have previously been shown to cause the familial hemiplegic migraine type 1 disorder (Bidaud et al., 2006).

The second region identified in this study of relevance for schizophrenia was the MHC, which is a multigenic region with high LD, a fact that impedes identification of disease-specific contributing genes. This region has also been identified as associated with both schizophrenia and bipolar disorder in previous large GWAS (Purcell et al., 2009; Shi et al., 2009; Stefansson et al., 2009). Similarly to the MHC, the third region to be identified in this study, MIR137, has also previously been suggested to play an important role in schizophrenia (Consortium, 2011). microRNAs are short (20-24 nucleotides) non-coding RNAs involved in post-transcriptional regulation of gene expression by affecting the stability and translation of mRNAs. MIR137 is involved in adult neurogenesis and neuronal maturation (Consortium, 2011). Finally, 60% of the identified loci contained lincRNAs, implicated in epigenetic mechanisms and development, which constitute the fourth identified pathway important for schizophrenia development (Ripke et al., 2013). Epigenetics refers to the regulation of DNA transcription without alteration of the original DNA sequence. It is controlled by DNA methylation and acetylation, histone modifications, and non-coding RNAs.

In addition to individual SNPs identified by GWAS CNVs also seem to play an important role in the schizophrenia disease mechanisms. Nine CNVs, both deletions and

duplications, have been identified in different regions of the genome (Sullivan et al., 2012). Several of these CNVs have also been identified in patients with autism, which suggests a common underlying neurodevelopmental pathway for these disorders. The identified structural changes consists of rare, but potent variants only identified in few individuals, and are thus only explaining the disease mechanism for a few independent cases (van Os and Kapur, 22; Sullivan et al., 2012).

Several environmental risk factors have also been associated with schizophrenia. High paternal age, season of birth, urban environment, as well as high and low birth weight are all such environmental risk factors that might contribute to the disease mechanisms (American Psychiatric Association, 2013; Byrne et al., 2003; Wegelius et al., 2013).

2.5 Migraine

Symptoms

Migraine is an episodic, neurological headache disorder with a complex pathophysiology. There are two main sub-types of the disorder: migraine without aura and migraine with aura, which differ in their clinical symptoms. Migraine without aura is characterized by severe headache attacks lasting between 4 and 72 hours. The headache is often unilateral and pulsating with a moderate or severe intensity and may be associated with nausea and/or photophobia and phonophobia. Migraine without aura is the most common subtype of migraine. At least five such attacks are required for a diagnosis. About one third of all migraine patients suffer from migraine with aura, which is characterized by focal neurological symptoms that usually precede or sometimes accompany the headache. The symptoms usually develop gradually over 5 to 20 minutes and last for less than 60 minutes. The visual symptoms include positive features (e.g., flickering lights, spots or lines) and/or negative features (i.e., loss of vision), sometimes accompanied by sensory symptoms and speech disturbances. The migraine with aura is usually, but not necessarily, associated with a headache attack with the features of migraine without aura. There is a female to male predominance with about 17% of females and 8% of males suffering from migraine in the European population (Stovner et al., 2006).

Genetics

Genetic research on familial hemiplegic migraine (FHM), a rare monogenic familial form of migraine, has been successful and identified three genetic variants in three different genes coding for neuronal ion channels (Pietrobon, 2007). These variants perturb the activity of the channels, resulting in increased neuronal excitability, which might explain the migraine pathophysiology. Despite the success of identifying the underlying genetic mechanisms for FHM, the genetic contribution to the more common forms of migraine has not yet been elucidated. The FHM form of migraine only constitutes a very small portion of all migraine patients, and it seems that the different forms do not share genetic mechanisms for causing the disorders (Schurks, 2012).

Until recently, genetic research on migraine has mainly focused on candidate genes and regions, such as pathways in the serotonergic and dopaminergic metabolisms, neurogenic inflammation, vascular function, and hormone regulation. So far the candidate gene approach has not yielded any robust results in identifying genes involved in migraine. Many studies showed no significant effects, and positive results were often not replicated. Although linkage studies have identified several chromosomal regions involved in migraine, the underlying genes involved in the disorder have not yet been identified.

Recent large-scale GWAS have, however, been more successful in identifying genes possibly involved in migraine disease mechanisms. An impressively large meta-analysis on migraine, including a total of 23,285 individuals with migraine and 95,425 population-matched controls, was recently published (Anttila et al., 2013). Twelve loci, five of them new, with genome-wide significance were identified to be associated with migraine ($p<5x10^{-8}$). Eight of the twelve identified loci are located in or immediately adjacent to genes with known function in synaptic or neuronal regulation, and some of them even exert regulation over one another. The single most significant SNP (p-value = $2.69x10^{-19}$), rs11172113, is located within the *LRP1* gene encoding an endocytic receptor involved in several cellular processes such as intracellular signaling, lipid homeostasis, and clearance of apoptotic cells. An eQTL analysis on brain tissue identified five additional genes that might be implicated in the disease mechanisms. Subgroup-specific analyses identified a larger number of significant loci for the migraine without aura group compared to the group with aura (6 versus 0 respectively), which was surprising considering the higher heritability estimates for migraine with aura (sibling recurrence risk for migraine with aura is 3.8 versus 1.9 for migraine without aura). The authors speculate that the observed difference could either be caused by a larger genetic heterogeneity within the migraine with aura group, or that rare variants with larger effect sizes are more important for migraine susceptibility for this subgroup. Despite the large sample size and the number of genome-wide significant SNPs (142 SNPs at 12 loci), no specific pathways could be identified, and a broad understanding of the disease mechanisms is still missing.

3 AIMS

The general aim of this thesis was to increase knowledge about genetic mechanisms involved in the disease etiology of four heritable complex disorders: bipolar disorder, major depression, schizophrenia, and migraine. To accomplish this overall goal, the specific aims addressed in the four papers were as follows:

Paper I

Investigate if polymorphisms in estrogen receptor (ER) DNA-binding sequences may be involved in the genetic mechanisms mediating mood disorders. A genome-wide study of ER DNA binding sequences in bipolar disorder and major depression patients and healthy controls was conducted. Due to reported gender differences in mood disorders, the association tests were performed in each gender separately.

Paper II

Identify rare genetic risk factors for bipolar disorder through genome-wide studies on highly penetrant CNVs in families with a high prevalence of the disease. The findings were followed up in a larger case-control sample of bipolar disorder and schizophrenia.

Paper III

Create a computational program to aid researchers in the process of selecting functional SNPs for association analyses in user-specified regions and/or motifs genome-wide. The program performance was validated through an example study on three different transcription factors in a material with bipolar disorder and major depression patients and controls.

Paper IV

Perform an association study in a Swedish population-based cohort of eight SNPs previously identified as risk factors for migraine in three recent GWAS in large European migraine cohorts.

4 MATERIALS

Patient materials for the studies included in this thesis were assembled from multiple large case-control and family-based study samples. In paper I and III we used materials from bipolar disorder and major depression patients and their matched controls. In the first stage of the analysis in paper II we studied pedigrees of patients with bipolar disorder and in the second stage we included more bipolar disorder patients and matched controls as well as schizophrenia or schizoaffective disorder patients and controls. For the study in paper IV we used genotype information from patients diagnosed with migraine and matched controls. Below follows a brief description of the different materials used.

4.1 The NIMH Genetics Initiative and case-control CNV data

The National Institute of Mental Health (NIMH) has funded a Human Genetics Initiative with the aim to store and distribute clinical data and biomaterials, including DNA samples and cell lines, to aid research on mental disorders. For bipolar disorder and schizophrenia, data collection began in 1991, and materials have been made available to qualified researchers since 1996. Psychiatric assessments were made by using a structured interview tool called "Diagnostic Interview for Genetic Studies (DIGS)," which provides additional data on family structure, age, sex, and other clinically relevant information. Pedigrees have been ascertained in which there are at least two affected individuals who are biologically related first-degree relatives.

Genotyping of the entire dataset was done by the Broad Institute Center for Genotyping and Analysis on the Affymetrix Genome-wide Human SNP Array 6.0 platform. For studies in paper I and III we requested genotype data for the 724,067 SNPs that survived the initial quality control steps.

The NIMH bipolar data has been collected in 4 separate waves (4 datasets with a total of 644 families). Parametric and non-parametric linkage analysis had been carried out on all waves of pedigrees, but no data had been made available regarding parametric scores in individual families. In order to identify families with the strongest genetic involvement in the disease etiology, for the study in paper II, we performed family-wise linkage analyses on pedigrees with available genome-wide microsatellite data. Based on the results from these analyses, we selected 277 individuals in 48 pedigrees for further studies. DNA samples were requested and sent for genotyping using the Illumina Human 610 quad chip at the SNP Technology Platform at Uppsala University, Sweden. Genome-wide SNP genotypes for 592,275 markers were successfully generated for 275 of the individuals.

In stage 2 of the analyses in paper II, we used additional large genotype data sets based on individuals affected with bipolar disorder, schizophrenia or schizoaffective disorder, and control samples. These materials were collected both through collaborations, by taking direct contact with authors of other studies on copy-number variation in psychiatric or other diseases, and by mining publicly available supplementary

information of original articles. A complete list of the different data sources is provided in Table 1 of paper II. Some of the studies used materials from overlapping sets of individuals, and great care was taken to remove duplicate samples. In total we assembled data from 3,683 bipolar cases, 7,242 schizophrenia or schizoaffective cases, and 16,747 controls.

4.2 The NESDA and NTR

The Netherlands Study of Depression and Anxiety (NESDA) is a longitudinal cohort study that started in 2004 and continued for 8 years. A total of 2,981 individuals with and without current symptoms, 18 to 65 years of age, were recruited mainly from different health care institutions all over the Netherlands to participate in the study. Individuals at different developmental stages of the diseases (first and recurrent episodes) were asked to participate, and the diagnosis was made based on the Composite International Diagnostic Interview (CIDI). In addition to participating in several extensive interviews, the study subjects also provided blood samples for studies on biomarkers, DNA, RNA, and gene-expression profiles. Saliva samples were collected at multiple time points in order to study stress systems. Some individuals even participated in functional MRI studies. DNA samples and questionnaires were also collected from family members when available.

The major depression control samples were derived from the Netherlands Twin Register (NTR), which was established in 1987 for the purpose of doing scientific research. Today the register includes 62,000 twins between birth and 18 years of age and 25,000 twins over 18 years.

Genotyping was performed on the Perlegen GWAS platform by Perlegen Science. 599,156 SNPs were genotyped and 435,291 of them survived the initial quality control steps. Genotype information derived from the database on Genotypes and Phenotypes (dbGaP) on major depression patients and their controls were used for the studies in paper I and III.

4.3 The Swedish Twin Register

The Swedish Twin Register at the Karolinska Institutet in Stockholm is the world's largest twin database with information on more than 85 000 twin pairs. The register was established in the 1960s and contains information on both monozygotic and dizygotic Swedish twins born between 1886 and 2000. This is a population-based cohort, and collected data include both questionnaires on health and medication for each individual and sometimes also blood samples for blood component analyses and DNA extraction.

For the studies in paper IV, we included individuals from the Swedish Twin Register born between 1935 and 1958 that in a health questionnaire had provided answers on questions regarding migraine headache and for whom genotype data was available. The genotyping was performed at the SNP&SEQ Technology Platform at Uppsala University using the

Illumina Human OmniExpress bead chip, which provided genotype information for about 730,000 SNPs per individual. Migraine was classified according to ICHD-II. There were 672 migraineurs and 5,822 non-migraineurs in our sample. We selected one individual from each twin pair; preferably keeping the twin with migraine when the diagnosis was discordant. After quality control and cryptic relatedness and population stratification analyses, we had 556 cases and 3,146 controls in our material.

5 METHODS

5.1 *In vitro* experiments

This section briefly describes the biochemical methods that were used to generate data from biological samples.

5.1.1 Chromatin Immunoprecipitation (ChIP)

The functional regions that our analyses in papers I and III are based on were identified by other groups using the ChIP technique, which is described in detail in the corresponding papers (Carroll et al., 2006; Lin et al., 2007; Zhao et al., 2010). It is one of the most popular techniques for the study of DNA-protein interactions, such as DNA-binding regions of transcription factors or histone modifications. In brief, the protein of interest (in our case the transcription factor) is added to a cell lysate where it binds to the chromatin. After cross-linking of the DNA-protein complex, the chromatin is fragmented and the lysate is cleared from cell debris. Immunoprecipitation is performed using antibodies that specifically bind to the protein of interest. After removing the non-bound chromatin, the cross-linking bounds are reversed and the protein is removed. The DNA-fragments are then purified and analyzed by polymerase chain reaction (PCR), DNA-microarrays (ChIP-on-ChIP), or direct DNA-sequencing.

One advantage of using transcription factor binding regions derived from ChIP-studies is that we have good evidence that we are looking at biologically active TF-binding sites. There are other places in the genome that might have the transcription factor motif, but lack binding sites for other critical anchor-proteins. These binding sites are probably not active, but it would be hard to sort them out if we solely relied on *in silico* research. Furthermore, several transcription factors do not bind to the DNA directly but via a complex with other transcription factors. Such sequences thus lack the TF motif, but since the ChIP technique captures all *in vivo* active binding regions, these are included as well. A second advantage of defining the TF-binding regions by ChIP-technique is therefore that functional regions without the TF motif also will be included.

5.1.2 Genome-wide SNP genotyping

During the last fifteen years there has been a rapid advancement in the development of high-throughput and cost-effective SNP genotyping methods. The microarray method has enabled genome-wide genotyping, which is the fundamental data-generating method for GWAS. SNPs are in linkage disequilibrium with other SNPs in specific haplotype blocks. By genotyping certain SNPs in these blocks scattered throughout the genome, genome-wide genotype information can be obtained. Modern platforms contain several hundred thousand and up to several millions of such SNPs. For each SNP, there is a short primer of DNA sequence attached to the chip. Fragmented DNA samples from the

individuals in the study are then added to the chip. The DNA hybridizes to the primers, which leads to a fluorescence signal. The intensity and color of the signal, as read by a scanner, determines the genotype (Ragoussis, 2009; Syvänen, 2005).

In papers I, III, and IV we used SNP genotype data for our disease-association analyses. In paper II we utilized the raw signal intensities of SNP genotype data to study the genome-wide distribution of CNVs in patients with bipolar disorder or schizophrenia and control individuals. CNV distribution was analyzed with the software package PennCNV, which uses a hidden Markov model to detect CNV signals from the SNP intensity signals on the genotyping platform (Wang et al., 2007).

5.2 *In silico* analyses

This section briefly describes the statistical and computational methods used to manage, analyze, visualize, and interpret the data.

5.2.1 Perl programing

Computer programing is an extremely useful tool when analyzing large amounts of data. Several useful toolboxes, such as PLINK, are available on the internet for general data analysis tasks, but without knowledge of computer programing one is restricted to only those features available in these toolboxes. By knowing a programing language, one gets far more independent in the possibilities of designing and analyzing data in different projects. In papers I, III and IV, we have been using the programming language Perl to aid us in our studies. Perl is an open-source software included in several operating systems. It was originally developed in 1987 by Larry Wall and has constantly been improved since then, with the latest stable version Perl 5.18 released in May 2013. The Perl programing language is very flexible and provides (among other things) easy and powerful processing of text files, which makes it extremely useful for genetic research. The recent advancements in genotyping and sequencing technologies have produced huge amount of data. The most efficient and least error-prone way of handling these large text files is by programing.

In paper I we used Perl programing extensively throughout the project. We used it to identify the overlapping regions between the three ChIP-studies providing DNA-binding regions for the ER transcription factor; we wrote a program that searched through the entire genome nucleotide by nucleotide for the transcription factor motif (which can be done amazingly fast with a computer program and is more or less impossible without it); we developed programs to identify SNPs placed in these regions and motifs by comparing the identified genomic location with information available in the HapMap data file; by writing additional code we could also merge our different candidate lists of SNPs and search for genotyped markers in available patient materials; we further wrote programs for identifying genotyped markers in linkage disequilibrium with markers that were not

genotyped, and finally merged the different output files so that in the end we had one list of genotyped SNPs in the regions and motifs we were interested in. This procedure would not have been possible by only using already available software tools.

Since we found this custom method useful and realized that, after some modifications, it could also be useful for others, we decided to merge all the independent scripts created during the work with paper I into one program. Thus in paper III we created a user-friendly software tool that can search for any genomic region and/or motif of interest genome-wide. We further extended the program by including several descriptive statistical analyses on the SNPs of interest.

5.2.2 *Linkage analysis*

In paper II we performed linkage analyses on the bipolar disorder material from NIMH in order to identify and select families with a high prevalence for the disease for our CNV analyses. Family-wise parametric and non-parametric linkage analyses were performed for the entire data set by using the program GENEHUNTER-PLUS (Kong and Cox, 1997). 48 families with 277 individuals were selected for further analyses.

5.2.3 *Quality control (QC)*

In addition to the initial quality control steps performed by the different genotyping companies, we performed independent quality control analyses on our materials mainly following standard quality control protocols for genetic case-control association studies (Clarke et al., 2011). In short: individuals and markers with a high degree of missing genotypes were excluded, as were non-informative SNPs with a minor allele frequency below 1 percent. We further excluded SNPs that in the control group showed a Hardy-Weinberg equilibrium (HWE) deviation of $p<0.0001$, which may indicate genotyping problems. The HWE principle says that allele and genotype frequencies in a population remain constant (they are in equilibrium) from generation to generation unless specific disturbing influences are introduced (such as mutation, selection, and non-random mating, which is the reason for only testing HWE in the control group). These quality control steps were performed by using the open source software toolbox PLINK (Purcell et al., 2007).

In paper IV, where we used a Swedish cohort of twins to study migraine, we performed cryptic- and subject-relatedness analyses to identify and exclude individuals with unknown relations (there could be multiple twin-pairs within the same family, which would inflate the p-values and give false positive signals). A population stratification analysis based on multidimensional scaling (MDS) calculations in four dimensions was also performed using PLINK. Outliers were excluded if they were deviating >6 standard deviations from the mean. A one-degree of freedom allelic Chi-square test was performed before and after all quality control steps in order to calculate the genomic inflation factor λ as a measure of the quality of the QC analyses.

5.2.4 Association analysis

For the association analyses in papers I and II we used PLINK and a two-sided Fisher exact test to calculate p-values, the odds ratio (OR), and 95% confidence intervals (CI) for SNPs placed in the transcription factor binding regions genome-wide. For the CNV-disease association analyses in paper II we used the statistical program R for our analyses (ref: R Core Team. R: A Language and Environment for Statistical Computing. [2.11.1]. 2012) A Fisher exact test with one degree of freedom was used to calculate p-values, OR, and 95% CI. For the association analyses on migraine in paper IV we used PLINK to perform logistic regression analyses using gender as a cofactor to eliminate any bias introduced by the predominance of female migraineurs in our material.

6 RESULTS AND DISCUSSION

6.1 Paper I: SNP in ER binding site increase risk for BD females

There is a female to male predominance in the prevalence of major depression, and depressive symptoms are more common in bipolar female patients compared to men with the same disorder (Bijl et al., 1998; Diflorio and Jones, 2010; Kessler et al., 2003; Weissman and Klerman, 1977). Involvement of the female sexual hormone estrogen has long been suggested as a possible explanation for the observed gender differences (Halbreich and Kahn, 2001; Payne, 2003). Previous studies had investigated the involvement of the estrogen receptor genes, but no one had looked at the sequences in the genome to which this transcription factor binds. Therefore we embarked on such a study. Genome-wide binding of the ER was investigated and SNPs identified in these regions were selected for following association studies. A flowchart illustrating the selection and inclusion process for the SNPs is shown in Figure 2 below.

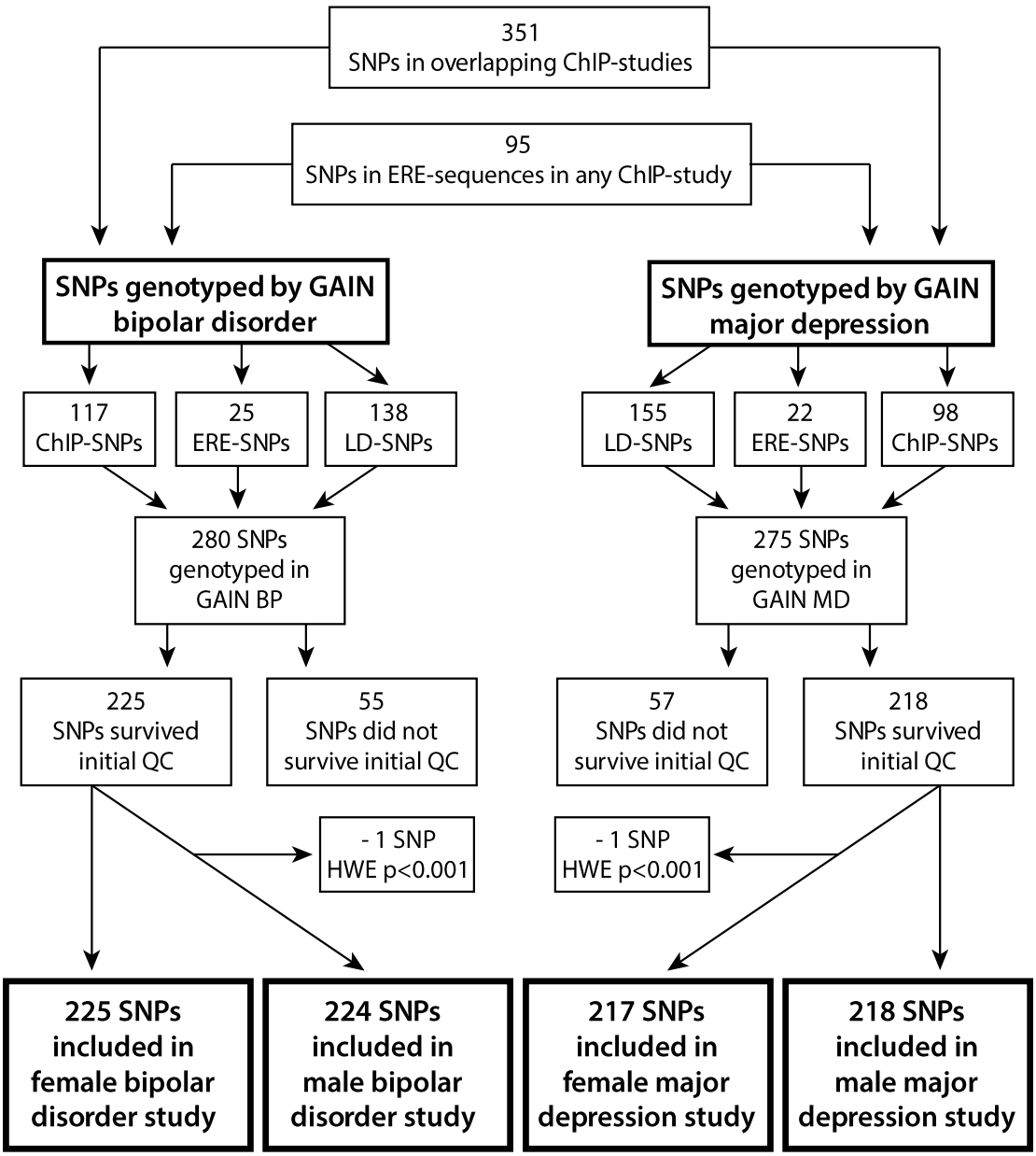

Figure 2 Flowchart illustrating the selection and inclusion process of the SNPs included in our study.

The materials used in this study are further described in the "Materials" section of this introduction and in papers I and III. Patients were diagnosed with either bipolar disorder or major depression, and association analyses were performed separately in each gender. A significant association was found for rs6023059 in females with bipolar disorder after correcting for multiple testing (corrected p-value = 0.023; OR 0.681; 95% CI 0.570-0.814). This SNP is located in the downstream region of the gene Transglutaminase 2 (*TGM2*), which has previously been shown to be implicated in neurodegenerative diseases (Bradford et al., 2009, 2011; Lesort et al., 2000). When bound to calcium, this protein performs cross-linking activities between various proteins in the cell and the activity has shown to be down-regulated by estrogen (Assisi et al., 1999; Fujita et al., 2006). This

indicates that fluctuating levels of estrogen might work as a triggering factor for bipolar disorder females with this specific genetic variant.

6.2 Paper II: *MAGI*-CNVs increase risk for BD and SZ

The study in paper II aimed at investigating the possible involvement of rare CNVs in the etiology of bipolar disorder and schizophrenia. The initial stage of the analyses only included bipolar cases, but due to reported genetic overlap between this disorder and schizophrenia, we decided to also include cases with a diagnosis of schizophrenia or schizoaffective disorder in later stages of the analyses in order to increase sample size and power. Stage 1 of our study was a family-based analysis of bipolar disorder families from the NIMH collection. In order to select individuals with a high prevalence of the disorder for further analyses we performed parametric and non-parametric linkage analyses in individual families. Based on these results we selected 277 individuals from 48 families for genome-wide genotyping. Genome-wide CNV distribution was evaluated from the genotype intensity data using PennCNV. The identified CNVs (which were not reported to be found in the general population) were ranked based on the number of affected individuals per family in which they were found. The highest-ranking CNV was a deletion (~200 kb) in intron 1 of the membrane-associated guanylate kinase 1 (*MAGI1)* gene found in six out of six affected individuals (see Fig 3 and 4). Further inspection of our pedigrees revealed another family with a duplication in the *MAGI* gene in two out of three affected individuals (see Fig 3 and 4).

In stage 2 of our analyses we sought additional support for this gene through analysis of a sample of 4084 cases with bipolar disorder, schizophrenia or schizoaffective disorder. Three CNVs (one deletion and two duplications) >100 kb were found in these cases (see fig 4). In stage 3 of our analyses we searched for further support of the involvement of this gene with the disorders by performing a meta-analysis of pooled data from our own as well as publicly available materials. 10,925 patients and 16,747 controls were included in the final data set. Three additional CNVs (one deletion and two duplications) were found in cases, and two CNVs (one deletion and one duplication) were identified in control subjects.

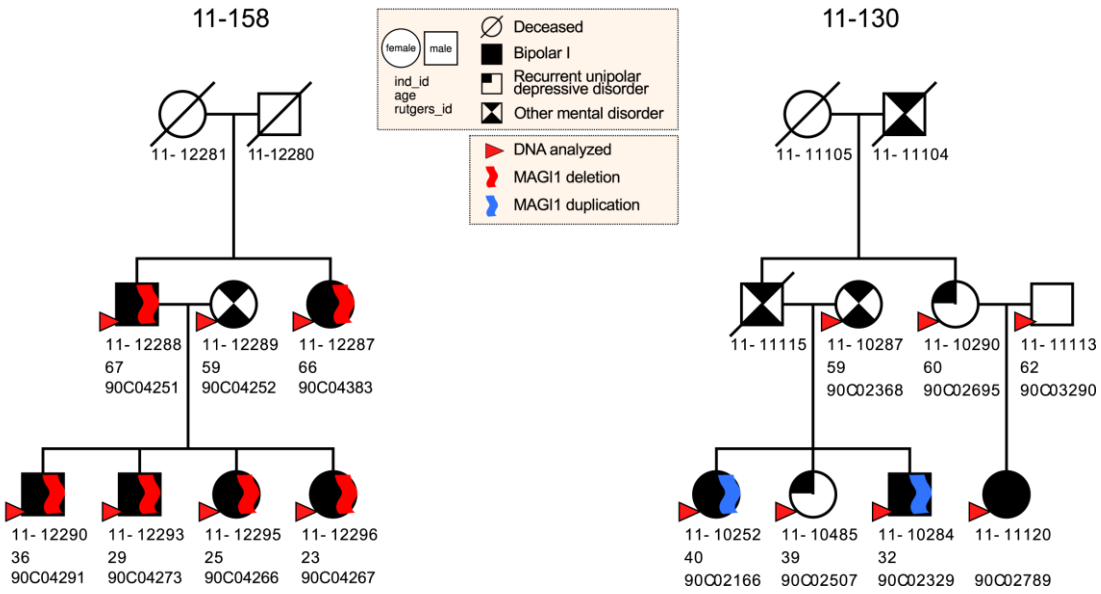

Figure 3 Pedigrees of familial bipolar disorder *MAGI*1 copy number variant carriers.

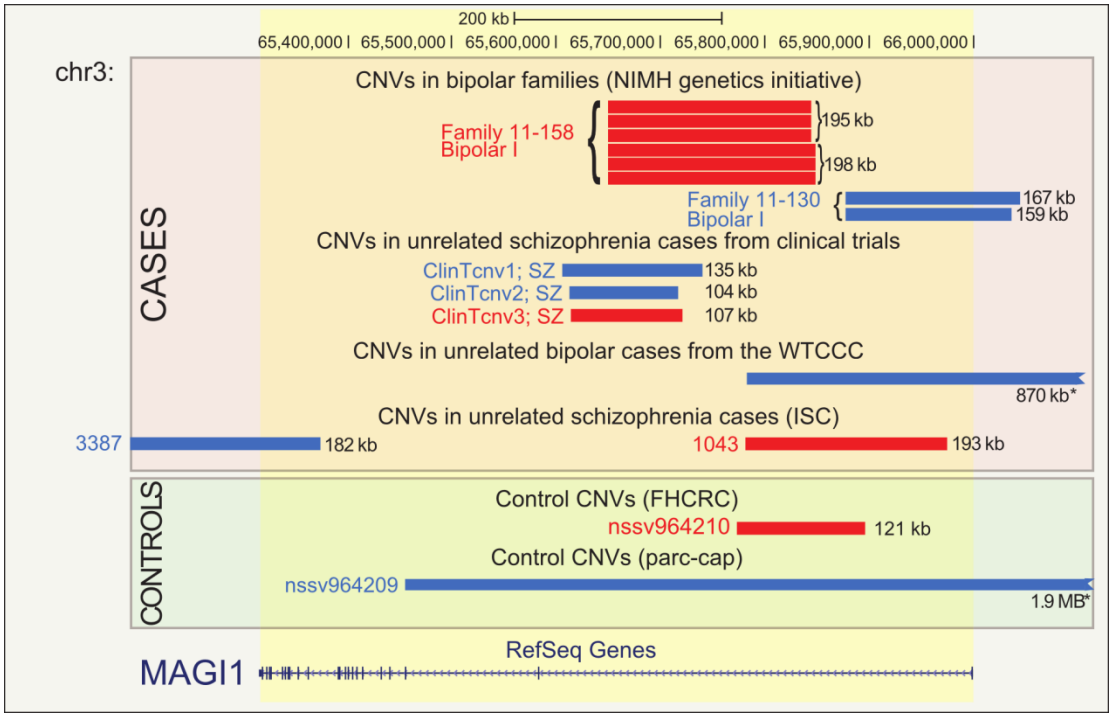

Figure 4 *MAGI*1 copy number variants (CNVs) over 100 kilobase (kb) identified in patients with bipolar disorder (BD), schizophrenia (SZ), schizoaffective disorder and controls. Red bars represent deletions, and blue bars duplications. *Marks a large CNV that extended past the genomic range shown in the figure. The genomic coordinates on top refers to the hg18 reference assembly of the human genome and chr3 = chromosome 3. The bottom line shows the extent of the coding (vertical lines), and non-coding (arrowheads) in the genomic sequence of the *MAGI*1 gene.

Due to previously reported associations between schizophrenia and the *MAGI2* gene, we extended our search to include *MAGI2* as well. This resulted in three identified CNVs (two deletions and one duplication) in schizophrenia cases and two duplications in control samples. In our association analysis on the pooled sample, the pedigree in which the variant was first discovered was not included, and the duplication appearing twice in another pedigree was only counted once. A Fisher's exact test with a one-sided hypothesis (alternative hypothesis: *MAGI1* CNV frequency higher in cases than in control subjects; null hypothesis: *MAGI1* CNV frequency lower in cases than control subjects or equal) was used for the association analysis. Our results (shown in Table 1) support a role for the involvement of rare CNVs in the *MAGI1* gene in a common disease pathway for bipolar disorder, schizophrenia and schizoaffective disorder.

Table 1 Association results from pooled analyses using one-sided Fisher's exact test with the hypothesis: *MAGI1* CNVs are more common in cases.

	Cases	Controls	Case CNVs	Control CNVs	p-value	OR [95% CI]
BD alone	3,683	16,747	2	2	0.15	4.53 [0.49, ∞]
BD, SZ, and SA	10,925	16,747	7	2	0.023	5.37 [1.26, ∞]

CNVs: copy number variations; OR: odds ratio; 95% CI: 95% confidence interval; BD: bipolar disorder; SZ: schizophrenia; SA: schizoaffective disorder

The *MAGI1* gene encodes a postsynaptic scaffolding protein that interacts with several molecules that previously have been suggested to be involved in the etiology of bipolar disorder and schizophrenia (e.g., beta-catenin, erythroblastic leukemia viral oncogene homolog 4 (ErbB4) [the neuregulin receptor], calcium channels [CACNA1], and glutamate receptors [α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor (AMPA)], see Figure 5). *CACNA1C* is one of the strongest findings in recent large GWAS for both bipolar disorder and schizophrenia (see this introduction under "Disorders"); ErbB4 and neuregulin have been associated with schizophrenia and bipolar disorder in numerous studies as has the AMPA receptor (see detailed discussion and references in paper II with supplemental information).

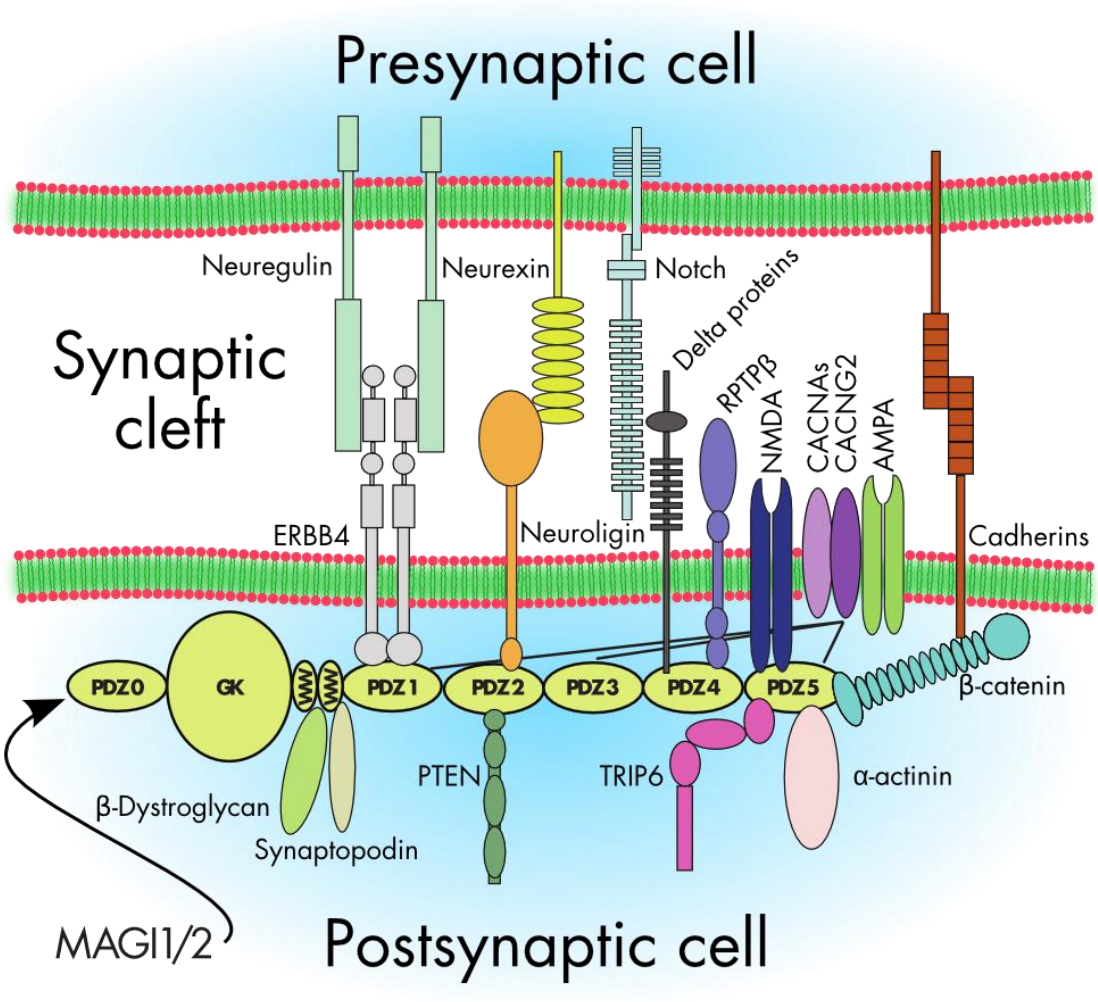

Figure 5 Cellular localization and functional interaction proteins for *MAGI1* and *MAGI2*.

6.3 Paper III: ReMo-SNPs identifies functional variants genome-wide

During the work with paper I, we created several independent Perl scripts to aid us in the inclusion and selection process of the SNPs. In paper II we merged these scripts together into one user-friendly program and added additional functions, so that the user could get descriptive statistical data on the analyses when needed. An overview of the input, action and output parts of the program is shown as a flowchart in Figure 6.

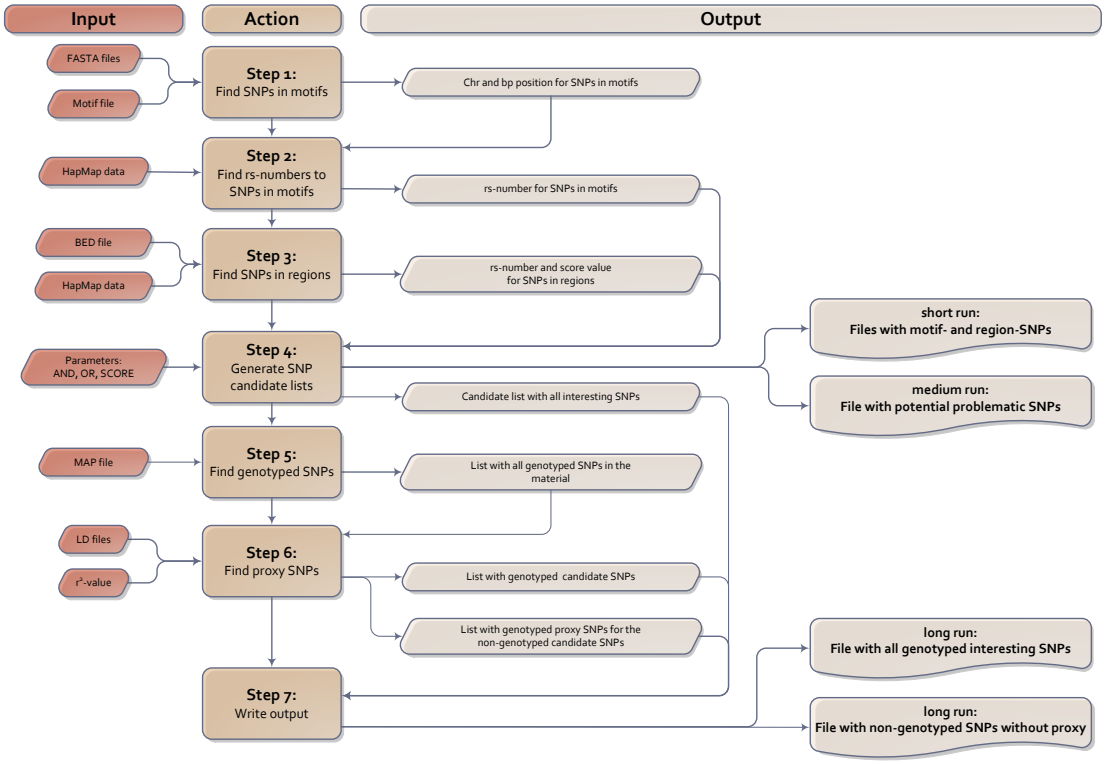

Figure 6 Flowchart illustrating an overview of the input, action and output parts of the ReMo-SNPs program (reproduced from paper III).

The ReMo-SNPs program is designed to identify SNPs in user-specified regions and/or motifs in order to aid the researcher in choosing SNPs with an increased likelihood of being functional for an association study. Embedded in the program is a feature to identify the selected markers in the user-provided list of markers. The program further finds proxy markers in high linkage disequilibrium according to a user-defined threshold for non-genotyped markers. The output files are thus ready to be used in a following association study.

In order to evaluate the power of the ReMo-SNPs program to identify functional SNPs we compared the SNPs identified by the program with three different functional software tools: Regulome (http://regulome.stanford.edu/), SNP Function Annotation Portal (http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx), and SNP Function Prediction (http://snpinfo.niehs.nih.gov/snpinfo/snpfunc.htm)

(Boyle et al., 2012; Wang et al., 2006; Xu and Taylor, 2009). For this evaluation we had two approaches; first we looked at motif-SNPs placed within versus outside of the user-specified regions; as a second approach we compared the scores obtained for the real motif-SNPs with control motif-SNPs within versus outside regions. With the first approach we found that in eight out of nine cases the motif-SNPs placed within regions yielded higher scores, which demonstrates the advantage of the possibility to combine regions and motifs in the search for functional SNPs. Results are shown in Figure 7. With the second approach the opposite pattern was found with control motifs outside regions generating the highest scores. We would have expected a random distribution among the control motifs, which may illustrate a difficulty in creating short and truly non-functional control motifs.

We also performed SNP density and distribution analyses. Similar to what has been reported in other studies, the ReMo-SNPs program found a higher SNP density in the transcription factor binding regions and even higher frequencies within the motif sequences compared to the average SNP density in the genome, see Figure 8 (Guo and Jamison, 2005). The non-random SNP distribution within the genome is well known and reflects the evolutionary pressure on different regions with, for example, less SNPs in exons compared to introns. Non-advantageous SNPs in coding regions have constantly been sorted out by natural selection, whereas in the genome in general they remain if they don't cause any deleterious effects. SNPs in regulatory regions may add an ability to adjust gene-regulation in a more fine-tuned way and may thus be most useful in the polymorphic state.

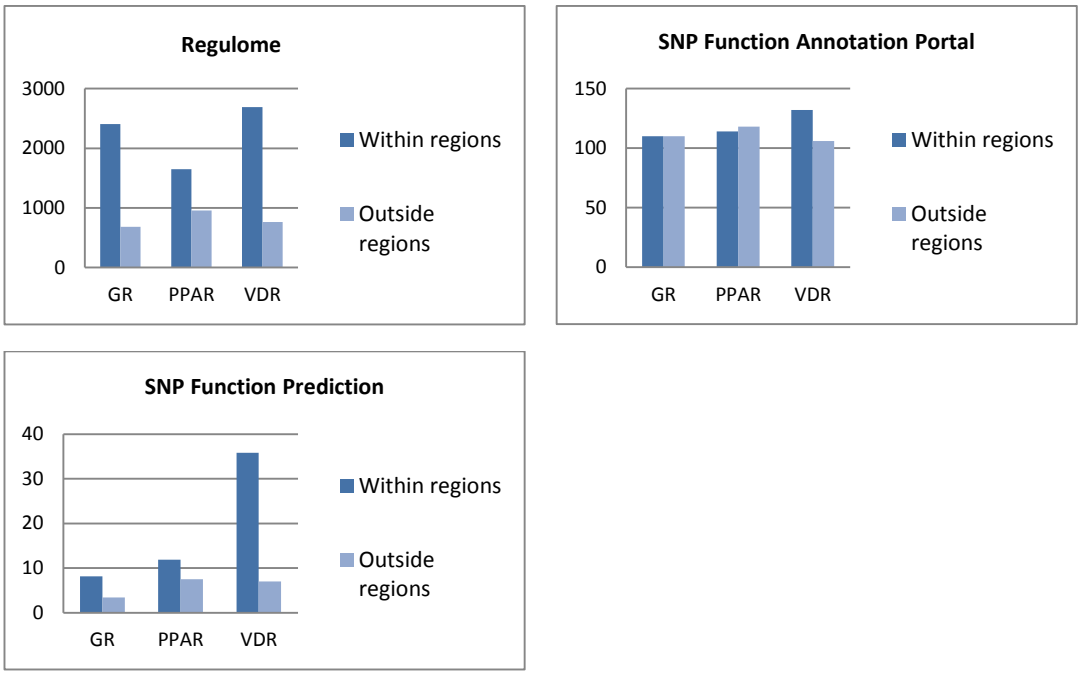

Figure 7 Original motif-SNPs placed within versus outside transcription factor binding regions. Score-values on the y-axes are unique for each program and should therefore not be compared between the different programs. GR: glucocorticoid receptor; PPAR: peroxisome proliferator-activated receptor; VDR: Vitamin D receptor.

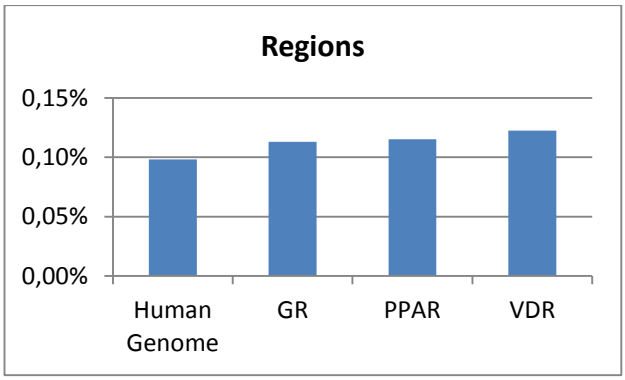

Figure 8 SNP densities in the transcription factor binding regions compared to the average SNP density in the human genome of the CEU population (CEU: Utah residents with Northern and Western European ancestry from the Centre d'Etude du Polymorphisme Humain (CEPH) collection). GR: glucocorticoid receptor; PPAR: peroxisome proliferator-activated receptor; VDR: Vitamin D receptor.

In SNP distribution analyses we further looked at the numbers of SNPs at different positions within the motifs, the number of SNPs at each nucleotide in the motif, and the number of SNPs per motif. Most motifs only had one SNP, and the SNP distribution for each type of SNP showed–as expected–that the G and C nucleotides harbored more SNPs compared to A and T. The analyses of SNP distributions within the motifs revealed a

complex pattern perhaps reflecting different mutation rates for different nucleotides in combination with different mutation likelihoods at different positions within the motif.

Association analyses of the three different transcription factors were performed on data from patients with bipolar disorder and major depression and controls. There was no particular a priori correlation of these TFs with the diseases, and they were chosen because of practical reasons and the availability of genome-wide experimentally validated binding data. No association remained significant after correcting for multiple testing.

6.4 Paper IV: Two SNPs are significantly associated with migraine

In paper IV, we performed an association study in a Swedish cohort on eight SNPs previously identified as genetic risk factors for migraine in three large European GWAS (Anttila et al., 2010; Chasman et al., 2011; Freilinger et al., 2012). Two of these SNPs, rs1835740 and rs2651899, were identified as risk factors for migraine also in the Swedish material after correcting for multiple testing, see Table 2. rs1835740 had a corrected p-value of 0.02704 (OR 1.38, 95% CI 1.128-1.686), and rs2651899 had a corrected p-value of 0.02816 (OR 1.24, 95% CI 1.082-1.1411). For both SNPs the minor alleles were more common in cases than in controls, which indicates that the minor alleles confer an increased risk for migraine.

Table 2 Results from association analysis on the eight SNPs included in the study using logistic regression with gender as cofactor.

SNP ID	Alleles	MAF cases	MAF controls	P-value	Corrected p-value	OR (95% CI)
rs2651899	C < T	0.4604	0.4035	0.00176	0.02816	1.24 (1.082- 1.411)
rs2274316	C < A	0.3732	0.3537	0.9414	1	1.01 (0.8615-1.174)
rs1003540	G < A	0.1862	0.2063	0.1623	1	0.82 (0.6281-1.081)
rs4075749	G < A	0.3318	0.3132	0.4464	1	1.06 (0.9102-1.238)
rs9349379	G < A	0.4406	0.4319	0.9920	1	1.0 (0.8695-1.148)
rs1835740	T < C	0.2239	0.1995	0.00169	0.02704	1.38 (1.128-1.686)
rs6478241	A < G	0.3687	0.3600	0.9228	1	1.01 (0.8717-1.164)
rs11172113	C < T	0.3804	0.4184	0.0845	1	0.89 (0.7744-1.016)

SNP = Single Nucleotide Polymorphism; alleles = minor allele < major allele; MAF = minor allele frequency; corrected p-value was calculated using the Bonferroni correction method; OR = odds ratio; 95% CI = 95% confidence interval.

The SNP-disease association was tested using logistic regression analysis with gender as cofactor to eliminate any bias introduced by the predominance of female migraineurs in our material. 556 cases (79% females) and 3,146 controls (48% females) were included in the final dataset. The material originated from the Swedish Twin Register, described under "Materials" in section 4. It is a population-based cohort, and the migraineurs included individuals both with and without aura.

To evaluate the distribution of carrying more than one risk allele, an unweighted cumulative risk analysis was performed for the eight SNPs included in our study. Risk alleles were defined using odds ratios published in a previous study (Anttila et al., 2010). Differences between cases and controls were evaluated using a Chi-square test for trend. The overall difference was significant with a Chi-square of 13.96 and p-value < 0.0002. The results are illustrated in Figure 9, showing that the migraineurs generally had a higher load of risk alleles compared with the control group in our material.

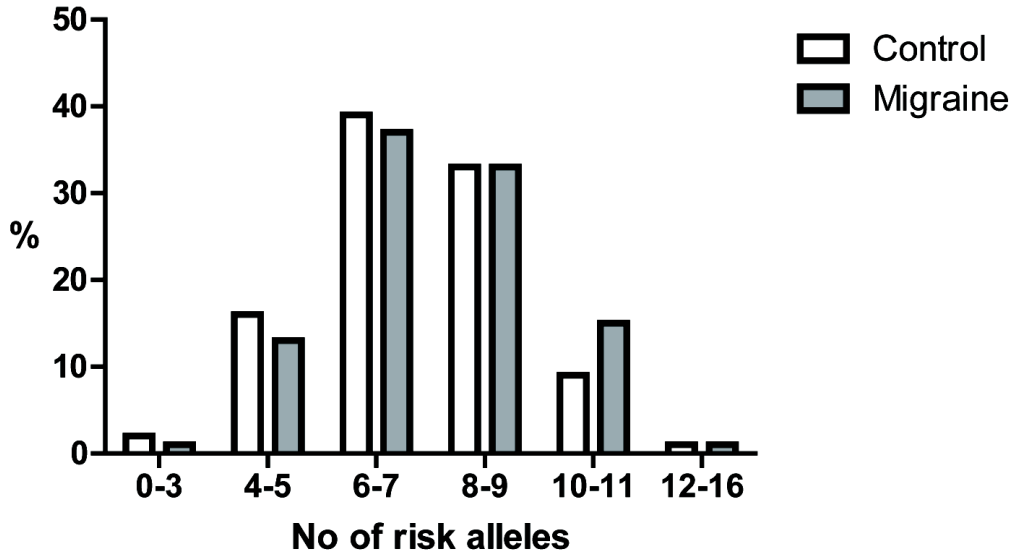

Figure 9 Results from the unweighted cumulative risk analysis showing differences in distributions of risk alleles between migraine patients and controls. Results were significant with a Chi-square of 13.96 and a p-value < 0.0002.

rs1835740 was first reported as a risk factor for migraine by Anttila et al. in a GWAS including seven clinical-based European cohorts. The cases were divided into subgroups based on a diagnosis of migraine with or without aura, and the effect of rs1835740 was stronger in migraineurs with aura. Depending on the cohort background, different forms of the disorder with variable severity may be represented in the different cohorts. In this study, the patients had actively searched medical assistance for their condition, which is why it might represent individuals with a more severe form of the disease compared to a population-based cohort. Moreover, it has been suggested that different phenotypes of migraine (e.g., with and without aura) might have different sets of genetic risk factors involved in the disease etiology. It was therefore interesting that rs1835740 was

significant also in the Swedish material, which is a population-based cohort including migraine patients both with and without aura. The other significant SNP, rs2651899, was first reported by Chasman et al. in a population-based cohort including migraine patients both with and without aura, which thus is in agreement with the background of the Swedish material.

One limitation of our study is that the diagnosis of migraine in our population was self-assessed based on a questionnaire using the ICHD-II criteria for migraine and not evaluated by a clinician.

7 CONCLUSIONS AND FUTURE PERSPECTIVE

7.1 Conclusions based on the studies in each one of the four papers

Paper I

This is the first study on ER-DNA binding variation in mood disorders. Our results support a gender difference in the etiology of bipolar disorder and suggest that differences in ER binding might contribute to an increased risk of developing bipolar disorder in females. Thus, fluctuating estrogen levels might work as a triggering factor for bipolar disorder in females with a specific genotype at rs6023059, close to the *TGM2* gene.

One limitation with this study was that the estrogen binding regions were specified based on Chromatin Immunoprecipitation studies in only one cell line. It would therefore be important to replicate these finding with binding information from other cell lines. In addition, replication in other bipolar disorder cohorts would also be of great importance to evaluate the role of the rs6023059 polymorphism in additional materials. Furthermore, functional studies need to be carried out to confirm the influence of the polymorphism on ER activity.

Paper II

Our results from the multistage analysis of rare CNV events in the *MAGI1* and *MAGI2* genes in bipolar disorder and schizophrenia patients and controls supports an involvement of these variants in the disease etiology. The biological function of the *MAGI1* and *MAGI2* genes as scaffolding proteins in the postsynaptic cell, with numerous interaction partners previously identified as associated with the disorders, further strengthens these findings.

One limitation of the study is that the DNA was derived from different sources (eg cell-lines and blood samples). DNA from the families where we first did the initial genome-wide scan for CNVs were derived from lympoblastoid cell lines, which increases the risk of the identified *MAGI1*-CNV being an artifact. However, the fact that the deletion was found in several individuals of the same family and with almost the same borders makes this highly unlikely. Another limitation is that different computational methods were used for detecting CNVs from the genotype data, which to some extent was overcome by the lower limit of 100 kb as CNV length.

Our study supports the hypothesis of the involvement of rare CNVs in the *MAGI1* and *MAGI2* genes as an underlying disease mechanism for bipolar disorder and schizophrenia. However, further studies in this region are needed in order to extend the knowledge on common variants in these genes. With a prior hypothesis and candidate-gene association studies, the multiple testing significance thresholds are not as high as for a GWAS, and further significant associations in these regions may be found in targeted

analyses. Furthermore, re-sequencing analyses in these genes would be highly interesting and perhaps reveal other types of variants in these genes that also could be important for the disease mechanisms. If CNVs in the *MAGI1* and/or *MAGI2* genes would be confirmed as highly penetrant disease susceptibility variants, they could perhaps be used as markers in diagnosis or genetic counseling in families with a high disease prevalence.

Paper III

ReMo-SNPs is a new computational tool to aid researchers in selecting functional SNPs in user-specified regions and/or motifs of interest genome-wide. It enables the researcher to combine *in silico* identified binding motifs with experimentally validated *in vivo* or *in vitro* binding regions. In addition, the program automatically identifies genotyped SNPs from the user-provided list of markers and replaces non-genotyped SNPs with SNPs in high linkage disequilibrium where available. This feature enables the user to directly use the generated output data in a following association study.

Even though we used transcription factor binding regions and motifs as input data in our example, this is not a limitation of the program. The ReMo-SNPs program is able to search in any type of region or motif genome-wide and is thus easy to use in several different genetic research areas, such as in the study of SNPs in binding regions of different cis-regulatory elements or DNase I hypersensitivity sites, or genomic regions with histone modifications or DNA methylations. Integration of *in silico* and *in vitro* data is likely to become increasingly important as our field struggles to translate ever-increasing amounts of data into information.

One limitation of ReMo-SNPs is that it is written in Perl, which is due to the fact that it constitutes a synthesis and extension of Perl scripts that we wrote for Paper I. It would run much faster if it were written in a compiled language (e.g., C#, Objective-C). The step that takes the longest time is to search through the entire genome for the motif of interest. Luckily, although our understanding of the genome has grown dramatically, the genome itself is of a fixed length. This means that ReMo-SNP will always be able to complete the analyses in a reasonable time frame, as encountered during our studies, which could take up to almost a day for a single transcription factor.

Paper IV

In our replication study on eight SNPs previously found to be associated with migraine in large European GWAS we found that two of the SNPs, rs1835740 and rs2651899, also were associated with migraine in Sweden. The rs1835740 risk allele is located between two genes involved in glutamate homeostasis: *MTDH* (metadherin) and *PGCP* (plasma glutamate carboxypeptidase). rs1835740 has previously been reported to be associated with higher *MTDH* expression (Anttila et al., 2010). The MTDH protein is involved in carcinogenesis and down-regulates *SLC1A2* (Solute carrier family 1, member 2), one of the major glutamate transporters in the brain. Downregulation of *SLC1A2* leads to

increased glutamate levels in the synaptic cleft, and *SLC1A2* knockout mice suffer from lethal spontaneous epileptic seizures.

rs2651899 is located in the first intron of the gene PR domain containing 16 (*PRDM16)*, the role of which in the underlying mechanisms of migraine has not been elucidated yet. It might turn out that the real disease-causing variant is not rs2651899, but another SNP in linkage disequilibrium with it. Another possibility could be that rs2651899 is linked with a copy number variant or another structural variation causing the disease. An interesting follow-up study on this marker would therefore be to re-sequence this region in a large number of cases carrying this risk allele in order to study this region more in detail. Another way to gain more information on the involvement of this region could be through fine-mapping SNPs located in the same haplotype block as rs2651899 in order to determine if there is another SNP with stronger association in this region. This could then be followed up by gene-expression studies. Furthermore, functional analyses on this region could also be done to evaluate if, for example, this region is involved in transcription factor binding, and if the variant causes disruption of such a site.

Unfortunately, we lack information from many of our patients regarding migraine subtype diagnosis (e.g., information regarding aura experiences). Otherwise, it would have been very interesting to perform studies on different subgroups of migraine to elucidate if they carry different sets of risk alleles. Another way to broaden the knowledge about genetic susceptibility of migraine in the Swedish population would be to increase the sample size by the use of an additional cohort. With an increase of sample size and/or a more narrow sub-diagnosis of migraine, other genetic risk factors involved in the migraine pathology might be identified as well. As always, further functional and genetic studies are needed in order to elucidate the underlying pathways and the genetic architecture for this disorder.

7.2 General conclusions based on the studies in the four papers

There are several aspects in the genetic analyses that are common and connect the four papers. First, in all four papers computer programing is used as a tool both to select markers and individuals, and to analyze and sort data and information. With the rapid advancements in genotyping, sequencing, and information technologies, vast amount of data are created in files well exceeding 1 million lines, which is why it has become almost essential to know at least one programming language to be able to work with those files in an efficient way. With knowledge on computer coding, work is not only far more efficient, but also much less error-prone. Language programing has made me more independent as a researcher and opened up my eyes for different ways of working with data and how to find quick and smart solutions for many problems. Coding skills will most probably be as important for researchers in genetics as is knowledge of anatomy for medical doctors.

Second, complex genetic disorders have, to some surprise, turned out to be – complex. Many researchers had hoped that once we would be able to make genome-wide analyses

we would quickly reveal a lot of the mystery behind the complex genetic diseases. Even though a lot of progress has been made and several new candidate genes and pathways have been identified, the big picture is still obscure. Therefore, in parallel with large efforts in mega- and meta-analyses of the whole genome, it is important to continue to carry out targeted and more hypothesis-driven studies on candidate genes and regions in order to identify the exact functional variants. Both the studies on different transcription factors in paper I and III and the study on CNVs (because they are large) in paper II are attempt to increase the a priori likelihood that they will identify functional variants that may contribute to the underlying disease mechanisms. The second most reliable evidence of finding a true genetic risk factor, after identifying a functional variant, is replication of the association, especially if it appears in other populations (Manolio, 2010). Paper IV represents such a study where we have performed, and found, associations of two previously identified risk variants for migraine that are of importance also for Swedish patients. Thus, this study strengthens the assumption of the contribution of these variants to the disorder and further increases knowledge about genetic mechanisms behind migraine in the Swedish population.

7.3 General conclusions and future perspective on complex genetics

7.3.1 Missing heritability

Given the substantial degree of heritability that has consistently been postulated for the psychiatric diseases, many researchers expected it to be relatively easy to identify the genetic contributions to the disease mechanisms once it became possible to examine the entire human genome. The lack of replicable and robust results in the field during the last decade has been very disappointing, and disease-causing genetic variants identified today only account for a very minor part of the heritability, a phenomenon which often is referred to as the "missing heritability". There has been a lively debate on the explanations of this phenomenon, where some propose it is due to the polygenicity of the disorders where many genes and genetic variants probably are not covered or identified by the currently used genotyping platforms or that the real disease-causing variants are not the ones identified, but are other variants in loose linkage disequilibrium with the identified ones.

Another possible explanation is that rare variants account for a much larger part of the disease incidence than we have expected, which could be an explanation for the lack of replication between studies. One limitation of GWAS is that they are designed to identify population variations in common markers (present in 5% or more of the population), which is why rare disease-causing variants will be undetected by this method (Manolio et al., 2009). Recent advancements in sequence technologies may, however, constitute a promising technique to fill this gap in the future. The traditional sequencing techniques have recently been replaced by rapid whole-genome or whole-exome parallel sequencing, so called next-generation sequencing. The throughput of sequence information has greatly increased at the same time as costs markedly have been reduced. With a

continuing technology development it may soon be possible to carry out studies with whole genome sequencing in many thousands of individuals. This will greatly facilitate the detection of rare variants and their contribution to disease mechanisms (Craddock and Sklar, 2013). Even though the latest SNP array platforms cover >1 million SNPs, there are many large parts of the genome that are still un-investigated and that may harbor some of the disease-causing variants. The next-generation whole genome sequencing may thus be very useful in detecting variants in these regions (Barnett and Smoller, 2009).

The expression of genes in our genome depends on interaction with many other genes that works as regulators on gene expression. This interaction between genes is called epistasis, and this may be a third explanation to the missing heritability. Thus, the effect by one variant may depend on the effect of several other variants that works in concert, which can make it impossible to detect any main effects by studying them separately (DeYoung and Clark, 2012).

Yet another explanation to the missing heritability concept is that the causative agent might be something that we either have not yet been able to detect with today's technology or that we just miss because we simply don't know what it is and how to look for it.

7.3.2 Environmental effects and triggering factors

In addition to genetic factors, environmental and internal effects also contribute to the disease mechanisms in complex genetic disorders. Internal and external factors can either work as triggering factors, inducing the onset of the disease or an episode of the disorder, or be part of the underlying disease mechanism. Stress, lack of sleep and hormonal fluctuations are known triggering factors for migraine patients that may increase the number of attacks (Mollaoğlu, 2013; Mulder et al., 2003). Urbanization and high parental age are environmental risk factors that seems to increase the risk for schizophrenia (Frans et al., 2011; van Os and Kapur, 22). It is not well known how the environmental risk factors influence the disease mechanisms. It might be that the combined genetic effects from many disease susceptibility loci make the system more vulnerable to environmental risk factors and internal triggering effects and therefore lower the threshold for disease liability. In some cases, environmental factors might also have a direct effect on the gene expression through modifications of epigenetic changes. It has been suggested that some environmental and internal factors, such as nutrition, stress, and pharmacological treatment might modify the epigenetic patterns and thereby influence gene expression (Menke et al., 2012).

7.3.3 Future strategies

There has been a lot of debate in the literature on different strategies on how to best design genetic studies in order to elucidate the missing heritability and identify the

disease-causing variants. One strategy is to combine different cohorts in order to increase sample size. The Psychiatric GWAS Consortium and the International Headache Genetics Consortium are two such large collaboration initiatives that aim at conducting meta-analyses of GWAS data. Larger sample sizes will obviously gain more power to detect common variants of smaller effect, and we have already started to see the benefits of such collaborations in diseases such as schizophrenia, bipolar disorder and migraine (Anttila et al., 2013; Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011; Ripke et al., 2013). In major depressive disorder, however a recent mega-analysis failed to detect disease causing variants despite an inclusion of 9,240 cases and 9,519 controls in the discovery phase of the study (Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium et al., 2013). This is probable due to several reasons, including the high prevalence in combination with relatively low heritability and high heterogeneity of the disorder and the fact that that the phenotypic predisposition between cases and controls is smaller compared to other psychiatric disorders such as schizophrenia (Sullivan et al., 2012; Wray et al., 2012).

One way to get around this could be to study a more homogenous group of cases with a more clearly identified phenotype or endophenotype. The underlying theory would then be that different genes may contribute to different phenotypes and a more phenotypically homogenous group of cases would probably have a more similar profile of risk alleles. For example, hospitalized major depressive cases, or those with a recurring and early onset variant may represent a more heritable phenotype. To increase the power even more one may also select individuals with a more homogenous environmental background (Wray et al., 2012).

A third promising strategy is to investigate genes or proteins in a network with shared biological pathways. One theory is that even though hundreds or perhaps thousands of common variants may be involved in the disease etiology they will be clustered in a far smaller number of genes, which will be involved in even fewer disease-causing pathways (Ripke et al., 2013; Sullivan et al., 2012). Small alterations in several genes involved in the same pathway may cause the system to become more sensitive to environmental changes. These changes therefore would have a more detrimental effect compared to a system lacking those gene alterations. If it turns out that the psychiatric disorders in fact are "pathway disorders" it may actually be of clinical benefit, since they would probably be both easier to diagnose and to treat pharmacologically (Sullivan et al., 2012). Typically GWAS identifies common variants with small effect sizes, nevertheless such variants may be of great importance pharmacologically if they lead to the identification of proteins that can become drug targets.

A majority of the genetic research today is done in Caucasian individuals, typically with a European background. However, in order to broaden our knowledge and identify susceptibility variants important for the global population it is utterly important to conduct genetic studies also in other populations. Phylogenetically older populations, such as populations in Africa, have a much more diverse genetic variability and shorter stretches of linkage disequilibrium in their genomes (Manolio, 2010; Manolio et al.,

2009). Therefore, genetic studies in these populations will probably not only give us more narrow localizations of the association signals, but also help us to identify more genetic susceptibility variants contributing to and explaining the underlying disease mechanisms for many complex genetic disorders.

8 ACKNOWLEDGEMENTS

The work of my PhD thesis was carried out at the Department of Neuroscience at Karolinska Institutet in Stockholm, Sweden. Many people have contributed to the work of this thesis in many ways and I sincerely thank everyone for the support and guidance I have received during this journey. I would like to especially thank the following people:

Andrea Carmine Belin, my main supervisor, for being an excellent scientist and for your amazing ability to organize and structure work. I could not have had a better guide throughout this last part of the thesis project. You have helped me to focus and to find strength to finish when I needed it most. Thanks for all your support on both the professional and personal level.

Silvia Paddock, my co-supervisor and former main supervisor, for endless enthusiasm and support throughout my journey. You are a brilliant scientist! Thanks for sharing your wisdom and incredible knowledge in the genetics field with me. Thanks for patiently and thoroughly answering all my questions and guiding my first steps into the world of science; for teaching me laboratory skills; for encouraging me to develop programing skills; and for helping me develop my scientific writing. I have learned so much from you!

Dagmar Galter, my co-supervisor, for being a fantastic scientist and always asking relevant questions, which markedly have helped me in my development as a scientist. Thanks for all the help with preparations for the histology teaching! Thanks also for your friendly company and for feeding me nuts and chocolate when I needed it most.

Mia (Maria) Lindskog, my co-supervisor, for all the fun discussions about science and life in general; for your amazing enthusiasm and warm friendship. I very much enjoyed our science as well as out-door conversations and exchanges of ideas and experiences in life!

Stefan Brené, former main supervisor, for teaching me animal laborative work and for letting me ask all these "stupid" questions without ever making me feel stupid.

Dai Wang, Reyna Favis, and Tomas Axelsson, co-authors on paper II, and **Patrik Magnusson and Nancy Pedersen,** co-authors on paper IV, for fruitful and pleasant collaborations and for critical reading of the articles.

Kerstin Iverfelt, my external mentor, for good advice when I needed it.

Lars Olson, for creating a creative and warm atmosphere at work and for sharing your vast knowledge and wisdom with us.

Robert Karlsson, co-author and former colleague, for your amazing ability to explain complicated statistical concepts and data programing and make it fully understandable; for all the good advice I constantly got from you throughout these years and for always being generous with your time to help me. Thanks for your friendship and fun company in the lab as well as on our journeys to conferences throughout the world.

Magnus Lekman, co-author and former colleague, for your warm friendship and many interesting discussions about science and life in general. I've learned a lot from you! Thanks for creating a pleasant atmosphere at work and for all the support you have given me throughout these years. I have very much enjoyed your company!

Anna Anvret, Sandra Gellhaar, Caroline Ran, and Sophia Savage, for all the fun discussions about science and life in general; for creating a very warm and friendly atmosphere at work; for your warm friendship and endless support and encouragement throughout these years.

Elin Åberg, Adam Sierakowiak, Anna Mattsson, Teresa Femenia Canto, Marta Gómez, Sahar Nikkhou Aski, and Peter Damberg, for enjoyable collaboration with the "Depression and exercise" manuscript. Elin and Teresa – it was fun working with you! Thanks for your warm friendship and for teaching me so much about animal studies! I have truly learned a lot from you! Thanks Adam for all the interesting discussions we have had! Thanks Anna for your support and friendship!

Mimi Westerlund, for your friendship and help with histology teaching preparations.

Linus Olson, for your warm friendship throughout these years.

All my past and present colleagues at the Olson lab, for creating a friendly and creative work environment. Thanks for all the interesting scientific discussions and discussions about life in general; and thanks for all support, help and encouragement I got from you throughout this journey! Thank you Anna Josephsson, Tobias Karlsson, Giuseppe Coppotelli, Jamie Ross, Jakob Kjell, Mathew Abrams, Alexandra Karlén, Astrid Björnebekk, Martin Werme, Matthias Erschbamer, Toshiki Endo, Fredrik Sterky Hansson, Jeff Blackinton, and all the Master's students who have made their way through our halls.

The past and present lab technicians in the Olson lab, Eva Lindqvist, Karin Pernold, Karin Lundströmer, Margareta Widing, and Katrin Wellfelt. Thank you all for your fantastic work and all your support! Eva – I've learned so much from you! Your knowledge and skills are amazing, thanks for all your help and support both in science and on a personal level.

Ida Engqvist, Filip Lindholm, and Roger Kjell for all the support and help with computer problems. You are invaluable!

Members of the Department of Neuroscience: the past and present heads of the department Staffan Cullheim and Sandra Ceccatelli, the administration staff and the staff at the animal department. Thanks for all the work you have done!

Special thanks also to **all the persons that have participated in the different cohorts** I've been using. Without you none of this would have been possible!

Louise Myhrman, my dear friend, who first evoked my interest for the fascinating world of neuroscience. Louise – you have changed my path in life in many ways and I am truly grateful for that. I have learned a lot from you and I am so thankful that I had the privilege to get to know you. You will always have a special place in my heart.

Caroline Lysell, Sophie Welin-Berger, Alva Vinterhed, and Nicce Törnberg, my closest friends, for always being there when I needed it most. You are truly wonderful!!

Ritva and Tapani Graae, my mother and father in law, for welcoming me into your family. Ritva – thanks for all your support! You are fantastic!

Gunnel and Bengt Hörnblad, my mother and father, for always believing in me, supporting and encouraging me. Thanks for everything!

Alvina and Isolde, my wonderful children, for giving me endless love and joy in life.

Christopher, the love of my life, thanks for everything! For your endless love and support on all levels. For sharing your life with me and making my life wonderful! Thanks for all the fantastic adventures we've had together! You are an amazing person and the best husband and father in the world!

Funding

This work has been supported by Karolinska Institutet KID-funding, Karolinska Institutet Funds, the Swedish Foundation for Strategic Research, the Swedish Research Council, the Swedish Brain Foundation, Swedish Brain Power and Åke Wibergs Stiftelse.

9 REFERENCES

American Psychiatric Association (2013). Diagnostic and Statistical Manual of Mental Disorders Fifth, Edition (Arlington, VA, American Psychiatric Association).

Anttila, V., Stefansson, H., Kallela, M., Todt, U., Terwindt, G.M., Calafato, M.S., Nyholt, D.R., Dimas, A.S., Freilinger, T., Müller-Myhsok, B., et al. (2010). Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. Nat. Genet. *42*, 869–873.

Anttila, V., Winsvold, B.S., Gormley, P., Kurth, T., Bettella, F., McMahon, G., Kallela, M., Malik, R., de Vries, B., Terwindt, G., et al. (2013). Genome-wide meta-analysis identifies new susceptibility loci for migraine. Nat. Genet. *45*, 912–917.

Assisi, L., Autuori, F., Botte, V., Farrace, M.G., and Piacentini, M. (1999). Hormonal control of "tissue" transglutaminase induction during programmed cell death in frog liver. Exp. Cell Res. *247*, 339–346.

Barnett, J.H., and Smoller, J.W. (2009). The genetics of bipolar disorder. Neuroscience *164*, 331–343.

Bidaud, I., Mezghrani, A., Swayne, L.A., Monteil, A., and Lory, P. (2006). Voltage-gated calcium channels in genetic diseases. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1763*, 1169–1174.

Bijl, R.V., Ravelli, A., and van Zessen, G. (1998). Prevalence of psychiatric disorder in the general population: results of The Netherlands Mental Health Survey and Incidence Study (NEMESIS). Soc. Psychiatry Psychiatr. Epidemiol. *33*, 587–595.

Blanchette, M. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. *16*, 656–668.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.

Bradford, M., Law, M.H., Stewart, A.D., Shaw, D.J., Megson, I.L., and Wei, J. (2009). The TGM2 gene is associated with schizophrenia in a British population. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. *150B*, 335–340.

Bradford, M., Law, M.H., Megson, I.L., and Wei, J. (2011). The functional significance of the TGM2 gene in schizophrenia: A correlation of SNPs and circulating IL-2 levels. J. Neuroimmunol. *232*, 5–7.

Bruder, C.E.G., Piotrowski, A., Gijsbers, A.A.C.J., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., et al. (2008). Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. Am. J. Hum. Genet. *82*, 763–771.

Burton, P.R., Tobin, M.D., and Hopper, J.L. (2005). Key concepts in genetic epidemiology. Lancet *366*, 941–951.

Byrne, M., Agerbo, E., Ewald, H., Eaton, W.W., and Mortensen, P.B. (2003). Parental age and risk of schizophrenia: a case-control study. Arch. Gen. Psychiatry *60*, 673–678.

Carroll, J.S., Meyer, C.A., Song, J., Li, W., Geistlinger, T.R., Eeckhoute, J., Brodsky, A.S., Keeton, E.K., Fertuck, K.C., Hall, G.F., et al. (2006). Genome-wide analysis of estrogen receptor binding sites. Nat. Genet. *38*, 1289–1297.

Chasman, D.I., Schürks, M., Anttila, V., de Vries, B., Schminke, U., Launer, L.J., Terwindt, G.M., van den Maagdenberg, A.M.J.M., Fendrich, K., Völzke, H., et al. (2011). Genome-wide association study reveals three susceptibility loci for common migraine in the general population. Nat. Genet. *43*, 695–698.

Cichon, S., Mühleisen, T.W., Degenhardt, F.A., Mattheisen, M., Miró, X., Strohmaier, J., Steffens, M., Meesters, C., Herms, S., Weingarten, M., et al. (2011). Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. Am. J. Hum. Genet. *88*, 372–381.

Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2011). Basic statistical analysis in genetic case-control studies. Nat. Protoc. *6*, 121–133.

Consortium, T.E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Consortium, T.S.P.G.-W.A.S. (GWAS) (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.

Cordell, H.J., and Clayton, D.G. (2005). Genetic association studies. Lancet *366*, 1121–1131.

Craddock, N., and Sklar, P. (2013). Genetics of bipolar disorder. Lancet *381*, 1654–1662.

Cross-Disorder Group of the Psychiatric Genomics Consortium, and Genetic Risk Outcome of Psychosis (GROUP) Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet *381*, 1371–1379.

Dawn Teare, M., and Barrett, J.H. (2005). Genetic linkage studies. Lancet *366*, 1036–1044.

DeYoung, C.G., and Clark, R. (2012). The gene in its natural habitat: The importance of gene–trait interactions. Dev. Psychopathol. *24*, 1307–1318.

Diflorio, A., and Jones, I. (2010). Is sex important? Gender differences in bipolar disorder. Int. Rev. Psychiatry *22*, 437–452.

Ferrari, A.J., Charlson, F.J., Norman, R.E., Flaxman, A.D., Patten, S.B., Vos, T., and Whiteford, H.A. (2013). The Epidemiological Modelling of Major Depressive Disorder: Application for the Global Burden of Disease Study 2010. PLoS ONE *8*, e69637.

Ferreira, M.A.R., O'Donovan, M.C., Meng, Y.A., Jones, I.R., Ruderfer, D.M., Jones, L., Fan, J., Kirov, G., Perlis, R.H., Green, E.K., et al. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat. Genet. *40*, 1056–1058.

Feuk L. (2012). Genomic structural variants: methods and protocols. (Humana; London: Springer, Totowa, N.J).

Frans, E.M., McGrath, J.J., Sandin, S., Lichtenstein, P., Reichenberg, A., Långström, N., and Hultman, C.M. (2011). Advanced paternal and grandpaternal age and schizophrenia: a three-generation perspective. Schizophr. Res. *133*, 120–124.

Freilinger, T., Anttila, V., de Vries, B., Malik, R., Kallela, M., Terwindt, G.M., Pozo-Rosich, P., Winsvold, B., Nyholt, D.R., van Oosterhout, W.P.J., et al. (2012). Genome-wide association analysis identifies susceptibility loci for migraine without aura. Nat. Genet. *44*, 777–782.

Fujita, K., Kato, T., Shibayama, K., Imada, H., Yamauchi, M., Yoshimoto, N., Miyachi, E., and Nagata, Y. (2006). Protective effect against 17beta-estradiol on neuronal apoptosis in hippocampus tissue following transient ischemia/recirculation in mongolian gerbils via down-regulation of tissue transglutaminase activity. Neurochem. Res. *31*, 1059–1068.

Green, E.K., Grozeva, D., Jones, I., Jones, L., Kirov, G., Caesar, S., Gordon-Smith, K., Fraser, C., Forty, L., Russell, E., et al. (2010). The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. Mol. Psychiatry *15*, 1016–1022.

Guo, Y., and Jamison, D.C. (2005). The distribution of SNPs in human gene regulatory regions. BMC Genomics *6*, 140.

Halbreich, U., and Kahn, L.S. (2001). Role of estrogen in the aetiology and treatment of mood disorders. CNS Drugs *15*, 797–817.

Headache Classification Committee of the International Headache Society (2013). The International Classification of Headache Disorders, 3rd edition (beta version). Cephalalgia *33*, 629–808.

Headache Classification Subcommittee of the International Headache Society (2004). The International Classification of Headache Disorders: 2nd edition. Cephalalgia Int. J. Headache *24 Suppl 1*, 9–160.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. *106*, 9362–9367.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.-C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998–1003.

Johansson, A.C.V., and Feuk, L. (2011). Characterization of copy number-stable regions in the human genome. Hum. Mutat. *32*, 947–955.

Jorde, L.B., Watkins, W.S., and Bamshad, M.J. (2001). Population genomics: a bridge from evolutionary history to genetic medicine. Hum. Mol. Genet. *10*, 2199–2207.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E., Wang, P.S., and National Comorbidity Survey Replication (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA J. Am. Med. Assoc. *289*, 3095–3105.

Kohli, M.A., Lucae, S., Saemann, P.G., Schmidt, M.V., Demirkan, A., Hek, K., Czamara, D., Alexander, M., Salyakina, D., Ripke, S., et al. (2011). The Neuronal Transporter Gene SLC6A15 Confers Risk to Major Depression. Neuron *70*, 252–265.

Kong, A., and Cox, N.J. (1997). Allele-Sharing Models: LOD Scores and Accurate Linkage Tests. Am. J. Hum. Genet. *61*, 1179–1188.

Lesort, M., Tucholski, J., Miller, M.L., and Johnson, G.V. (2000). Tissue transglutaminase: a possible role in neurodegenerative diseases. Prog. Neurobiol. *61*, 439–463.

Lewis, C.M., Ng, M.Y., Butler, A.W., Cohen-Woods, S., Uher, R., Pirlo, K., Weale, M.E., Schosser, A., Paredes, U.M., Rivera, M., et al. (2010). Genome-wide association study of major recurrent depression in the U.K. population. Am. J. Psychiatry *167*, 949–957.

Lichtenstein, P., Björk, C., Hultman, C.M., Scolnick, E., Sklar, P., and Sullivan, P.F. (2006). Recurrence risks for schizophrenia in a Swedish national cohort. Psychol. Med. *36*, 1417–1425.

Lin, C.-Y., Vega, V.B., Thomsen, J.S., Zhang, T., Kong, S.L., Xie, M., Chiu, K.P., Lipovich, L., Barnett, D.H., Stossi, F., et al. (2007). Whole-genome cartography of estrogen receptor alpha binding sites. PLoS Genet. *3*, e87.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2013). The database of genomic variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. gkt958.

Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne, E.M., Blackwood, D.H.R., Boomsma, D.I., et al. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. Mol. Psychiatry *18*, 497–511.

Malhotra, D., McCarthy, S., Michaelson, J.J., Vacic, V., Burdick, K.E., Yoon, S., Cichon, S., Corvin, A., Gary, S., Gershon, E.S., et al. (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron *72*, 951–963.

Manolio, T.A. (2010). Genomewide Association Studies and Assessment of the Risk of Disease. N. Engl. J. Med. *363*, 166–176.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

McGrath, J., Saha, S., Chant, D., and Welham, J. (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiol. Rev. *30*, 67–76.

Menke, A., Klengel, T., and Binder, E.B. (2012). Epigenetics, depression and antidepressant treatment. Curr. Pharm. Des. *18*, 5879–5889.

Mollaoğlu, M. (2013). Trigger factors in migraine patients. J. Health Psychol. *18*, 984–994.

Muglia, P., Tozzi, F., Galwey, N.W., Francks, C., Upmanyu, R., Kong, X.Q., Antoniades, A., Domenici, E., Perry, J., Rothen, S., et al. (2010). Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. Mol. Psychiatry *15*, 589–601.

Mulder, E.J., Van Baal, C., Gaist, D., Kallela, M., Kaprio, J., Svensson, D.A., Nyholt, D.R., Martin, N.G., MacGregor, A.J., Cherkas, L.F., et al. (2003). Genetic and environmental influences on migraine: a twin study across six countries. Twin Res. Off. J. Int. Soc. Twin Stud. *6*, 422–431.

Van Os, J., and Kapur, S. (22). Schizophrenia. The Lancet *374*, 635–645.

Palmer, L.J., and Cardon, L.R. (2005). Shaking the tree: mapping complex disease genes with linkage disequilibrium. Lancet *366*, 1223–1234.

Payne, J.L. (2003). The role of estrogen in mood disorders in women. Int. Rev. Psychiatry Abingdon Engl. *15*, 280–290.

Pietrobon, D. (2007). Familial hemiplegic migraine. Neurotherapeutics *4*, 274–284.

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. *43*, 977–983.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P., (Leader), S.M.P., Ruderfer, D.M., McQuillin, A., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

Ragoussis, J. (2009). Genotyping Technologies for Genetic Research. Annu. Rev. Genomics Hum. Genet. *10*, 117–133.

Rietschel, M., Mattheisen, M., Frank, J., Treutlein, J., Degenhardt, F., Breuer, R., Steffens, M., Mier, D., Esslinger, C., Walter, H., et al. (2010). Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. Biol. Psychiatry *68*, 578–585.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. *45*, 1150–1159.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.

Rucker, J.J.H., Breen, G., Pinto, D., Pedroso, I., Lewis, C.M., Cohen-Woods, S., Uher, R., Schosser, A., Rivera, M., Aitchison, K.J., et al. (2013). Genome-wide association analysis of copy number variation in recurrent depressive disorder. Mol. Psychiatry *18*, 183–189.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.

Schurks, M. (2012). Genetics of migraine in the age of genome-wide association studies. J. Headache Pain *13*, 1–9.

Sherry, S.T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P.A., Whittemore, A.S., Mowry, B.J., et al. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature *460*, 753–757.

Shi, J., Potash, J.B., Knowles, J.A., Weissman, M.M., Coryell, W., Scheftner, W.A., Lawson, W.B., DePaulo, J.R., Jr, Gejman, P.V., Sanders, A.R., et al. (2011). Genome-wide association study of recurrent early-onset major depressive disorder. Mol. Psychiatry *16*, 193–201.

Shyn, S.I., Shi, J., Kraft, J.B., Potash, J.B., Knowles, J.A., Weissman, M.M., Garriock, H.A., Yokoyama, J.S., McGrath, P.J., Peters, E.J., et al. (2011). Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. Mol. Psychiatry *16*, 202–215.

Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietiläinen, O.P.H., Mors, O., Mortensen, P.B., et al. (2009). Common variants conferring risk of schizophrenia. Nature *460*, 744–747.

Stovner, L.J., Zwart, J.-A., Hagen, K., Terwindt, G.M., and Pascual, J. (2006). Epidemiology of headache in Europe. Eur. J. Neurol. Off. J. Eur. Fed. Neurol. Soc. *13*, 333–345.

Sullivan, P.F., de Geus, E.J.C., Willemsen, G., James, M.R., Smit, J.H., Zandbelt, T., Arolt, V., Baune, B.T., Blackwood, D., Cichon, S., et al. (2009). Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. Mol. Psychiatry *14*, 359–375.

Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. *13*, 537–551.

Syvänen, A.-C. (2005). Toward genome-wide SNP genotyping. Nat. Genet. *37*, S5–S10.

The ENCODE Project Consortium (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. *9*, e1001046.

Wain, L.V., Armour, J.A.L., and Tobin, M.D. (2009). Genomic copy number variation, human health, and disease. Lancet *374*, 340–350.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.

Wang, P., Dai, M., Xuan, W., McEachin, R.C., Jackson, A.U., Scott, L.J., Athey, B., Watson, S.J., and Meng, F. (2006). SNP Function Portal: a web database for exploring the function implication of SNP alleles. Bioinformatics *22*, e523–e529.

Wegelius, A., Pankakoski, M., Lehto, U., Suokas, J., Häkkinen, L., Tuulio-Henriksson, A., Lönnqvist, J., Paunio, T., and Suvisaari, J. (2013). An association between both low and high birth weight and increased disorganized and negative symptom severity in schizophrenia and other psychoses. Psychiatry Res. *205*, 18–24.

Weissman, M.M., and Klerman, G.L. (1977). Sex differences and the epidemiology of depression. Arch. Gen. Psychiatry *34*, 98–111.

World Health Organization (WHO) (1994). International Classification of Diseases (ICD-10).

Wray, N.R., Pergadia, M.L., Blackwood, D.H.R., Penninx, B.W.J.H., Gordon, S.D., Nyholt, D.R., Ripke, S., MacIntyre, D.J., McGhee, K.A., Maclean, A.W., et al. (2012). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. Mol. Psychiatry *17*, 36–48.

Xu, Z., and Taylor, J.A. (2009). SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res. *37*, W600–W605.

Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., et al. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. *13*, R48.

Zhao, C., Gao, H., Liu, Y., Papoutsi, Z., Jaffrey, S., Gustafsson, J.-A., and Dahlman-Wright, K. (2010). Genome-wide mapping of estrogen receptor-beta-binding regions reveals extensive cross-talk with transcription factor activator protein-1. Cancer Res. *70*, 5174–5183.